PENYEIMBANGAN DATA PADA KLASIFIKASI DENGAN *SUPPORT VECTOR MACHINE* TERHADAP DATA PEMBAYARAN PINJAMAN BANK

Delvin Wang[1*], Paulina Heruningsih Prima Rosa[2]
[1,2] Universitas Sanata Dharma, [*]Penulis Koresponden
e-mail:[1]delvinwang876@gmail.com,[2] rosa@usd.ac.id

ABSTRACT

Loan default in banking can cause losses. Therefore, lenders need to predict the criteria of customers who fail to pay their loans. In this study, a classification model was built to predict customers who fail to pay bank loans by applying the Support Vector Machine algorithm, especially with the Radial Basis Function (RBF) kernel. Because the occurrence of default is not balanced with the occurrence of smooth payments, a data balancing process was carried out. This study also compared the effect of data balancing methods using Random Over Sampling and Near Miss techniques on the performance of the SVM algorithm. The dataset used is the Loan Default Prediction Dataset taken from the kaggle site, which consists of 255,347 records and 18 attributes. The results showed that the SVM model trained without data balancing had the highest accuracy of 88.49%, but with a recall of only 10% and an F1-score of 17%. After using ROS, the model accuracy decreased slightly to 83.52%, but the recall increased significantly to 94% and the F1-score to 89%. With Near Miss, the model accuracy drops further by 65.29%, but produces better precision and recall compared to without balancing. It can be concluded that balancing with ROS provides the best performance in terms of the balance between precision and recall, as seen from the highest F1-score value among the three methods.

Keywords: *Balancing, Loan Default, Near Miss, Radial Basis Function, Random Over Sampling, Support Vector Machine*

INTISARI

Kegagalan pembayaran pinjaman di dunia perbankan dapat menyebabkan kerugian. Oleh karena itu maka pemberi pinjaman perlu memprediksi kriteria nasabah yang gagal dalam membayar pinjaman. Dalam penelitian ini dibangun model klasifikasi guna memprediksi nasabah yang gagal membayar pinjaman bank dengan menerapkan algoritma Support Vector Machine, khususnya dengan kernel Radial Basis Function (RBF). Karena kejadian gagal bayar tidak seimbang dengan kejadian yang pembayarannya lancar, maka dilakukan proses penyeimbangan data. Dalam penelitian ini juga dibandingkan pengaruh metode penyeimbangan data memakai teknik Random Over Sampling dan Near Miss terhadap kinerja algoritma SVM. Dataset yang digunakan adalah Loan Default Prediction Dataset yang diambil dari situs kaggle, yang terdiri dari 255.347 record dan 18 atribut. Hasil penelitian menunjukkan bahwa model SVM yang dilatih tanpa penyeimbangan data memiliki akurasi tertinggi 88.49%, namun dengan recall hanya sebesar 10% dan F1-score 17%. Setelah menggunakan ROS, akurasi model sedikit menurun menjadi 83.52%, namun recall meningkat signifikan hingga 94% dan F1-score menjadi 89%. Dengan Near Miss, akurasi model menurun lebih jauh 65.29%, namun menghasilkan presisi dan recall yang masih lebih baik dibandingkan tanpa penyeimbangan. Dapat disimpulkan bahwa penyeimbangan dengan ROS memberikan kinerja terbaik dalam hal keseimbangan antara presisi dan recall, yang terlihat dari nilai F1-score tertinggi di antara ketiga metode.

Kata kunci: Gagal Bayar Pinjaman, Near Miss, Penyeimbangan data, Radial Basis Function, Random Over Sampling, Support Vector Machine

1. PENDAHULUAN

Peminjaman kredit adalah transaksi yang disepakati antara dua pihak dimana salah satu pihak meminjamkan uang atau dana kepada seseorang atau badan usaha agar orang atau badan usaha tersebut dapat menjalankan perusahaannya, mengejar tujuan atau kegiatan lainnya (Sembiring et al., 2022). Agar pinjaman disetujui, nasabah harus mengajukan permohonan terlebih dahulu, dan kemudian perusahaan pembiayaan harus memastikan bahwa konsumen memenuhi syarat. Formulir yang harus diisi oleh pemohon berisi informasi spesifik seperti status pernikahan, usia, tingkat pendidikan, pendapatan, jumlah pinjaman, dan lain sebagainya (Pahlevi et al., 2023).

Karena banyaknya nasabah melakukan pinjaman dan risiko gagal membayar pinjaman dapat menyebabkan kerugian, maka pihak yang memberikan pinjaman seperti bank perlu memprediksi kriteria nasabah yang gagal dalam membayar pinjaman. Penambangan data, khususnya metode klasifikasi dapat dimanfaatkan untuk mengatasi persoalan tersebut. Klasifikasi akan dilakukan berdasarkan parameter – parameter untuk menemukan pola dalam kumpulan data yang besar (Nalatissifa et al., 2021). Ketidakseimbangan pada dataset dapat mempengaruhi hasil klasifikasi, misalnya model menjadi bias terhadap kelas mayoritas atau model cenderung memprediksi kelas mayoritas lebih baik daripada kelas minoritas. Oleh karena itu, perlu dilakukan proses penyeimbangan data (*data balancing*) pada dataset yang tidak seimbang.

Beberapa peneliti sebelumnya telah melakukan klasifikasi dengan penyeimbangan data pada dataset yang tidak seimbang. Mallidi & Zagabathuni melakukan klasifikasi untuk analisis deteksi penipuan kartu kredit menggunakan Random Forest dan SMOTE untuk penyeimbangan data, menghasilkan akurasi terbaik sebesar 99.95% (Mallidi & Zagabathuni, 2021). Penelitian untuk memprediksi reaksi berdasarkan transaksi pelanggan menggunakan algoritma SVM kernel RBF dan SMOTE untuk penyeimbangan data memperoleh hasil akurasi 89% menggunakan parameter C = 1 dan *gamma* = 0.1 (Hayder et al., 2023). Tamami & Kharisudin melakukan komparasi metode SVM dan Naïve Bayes untuk pemodelan kualitas pengajuan kredit dan menerapkan teknik penyeimbangan data *over sampling* dan *under sampling*, dengan kesimpulan bahwa algoritma SVM dengan kernel Gaussian Radial Basis Function (RBF) menghasilkan akurasi tertinggi 96.65% dengan parameter yang digunakan C = 5 dan *gamma* = 2 (Tamami & Kharisudin, 2023). Pada penelitian deteksi penipuan kartu kredit di bawah data yang sangat tidak seimbang menunjukkan bahwa teknik *Random Oversampling* (ROS) yang digunakan bersama algoritma SVM memberikan akurasi yang signifikan, yaitu mencapai 94.7%, Presisi 98%, *Recall* 91.3%, *F1-Score* 94.5% (Singh et al., 2022). Penelitian prediksi gagal bayar kartu kredit pada dataset tidak seimbang menggunakan Gradient Boosted Decision Tree dan Near Miss sebagai penyeimbang data, mendapatkan akurasi sebesar 82.8%, Presisi, 80.5%, dan *Recall* 87% (Alam et al., 2020). Penelitian tentang analisis mendalam teknik *oversampling* untuk dataset tidak seimbang dengan menggunakan algoritma klasifikasi *Random Forest* dan teknik penyeimbang data *Random Oversampling* menghasilkan akurasi sebesar 99.7%, Presisi 93.4%, *Recall* 84.5% dan *F1-Score* 88.7% (Wibowo & Fatichah, 2021). Penelitian mengenai survei anomali kartu kredit dan deteksi penipuan menggunakan teknik sampling menggunakan 3 algoritma klasifikasi seperti *Random Forest*, *Logistic Regression* dan KNN dengan beberapa teknik sampling salah satunya *Random Oversampling* mendapatkan akurasi sebesar 99.97% (Alamri & Ykhlef, 2022).

Banyak teknik penyeimbangan data yang dapat diterapkan pada dataset yang tidak seimbang. Penelitian ini bertujuan untuk menerapkan algoritma SVM untuk memprediksi nasabah yang gagal membayar pinjaman bank dengan menerapkan algoritma Support Vector Machine (SVM), khususnya dengan kernel Radial Basis Function (RBF). Selain itu, juga untuk melihat pengaruh penyeimbangan data menggunakan *Random Over Sampling* dan *Near Miss* terhadap hasil akurasi algoritma SVM untuk mengklasifikasikan kriteria nasabah yang gagal dalam membayar pinjaman bank. Dataset yang digunakan adalah dataset *Loan Default Prediction Dataset* yang diambil dari situs Kaggle.

Support Vector Machine adalah teknik pembelajaran mesin yang biasanya digunakan untuk masalah klasifikasi dan regresi (Burges, 1998). Pada penelitian ini digunakan *Support Vector Machine* non-linier yang merupakan variasi dari SVM linier yang memungkinkan penanganan pola-pola yang tidak dapat dipisahkan secara linier dengan menggunakan teknik pemetaan ke dalam ruang berdimensi tinggi dengan menggunakan fungsi kernel. Adapun kernel yang digunakan pada penelitian ini adalah Kernel *Radial Basis Function* (RBF). Kernel RBF merupakan kernel yang dapat menangani kasus non-linier dengan mentransformasikan data ke dimensi yang tak terbatas (*infinite*). Transformasi ini dilakukan menggunakan fungsi *gaussian* sehingga distribusi data menjadi pola yang memungkinkan untuk dipisahkan oleh *hyperplane* (Burges, 1998).

Random Over Sampling (ROS) adalah pendekatan yang digunakan untuk mengurangi ketidakseimbangan antara kelas mayoritas dan minoritas. Dalam *Random Over Sampling*, sampel kelas minoritas ditambah dengan menggandakan data dan ditambahkan ke kelas minoritas (Najadat et al., 2020). Sedangkan *Near Miss* adalah salah satu teknik pengurangan jumlah data mayoritas yang digunakan dalam konteks klasifikasi ketika kita memiliki ketidakseimbangan kelas. Metode ini memilih sampel dari kelas mayoritas yang memiliki jarak terdekat dengan sampel dari kelas minoritas lalu algoritma mengurangi ukuran kelas mayoritas (Botchey et al., 2020).

2. METODE PENELITIAN

Pada penelitian ini digunakan model klasifikasi *Support Vector Machine* untuk mengklasifikasikan data tentang *Loan Default Prediction* dan juga untuk melihat pengaruh dari penyeimbangan data (*data balancing*)

menggunakan metode *Random Over Sampling* dan *Near Miss*. Pembangunan model klasifikasi dan pengujiannya dilakukan dengan menggunakan Python sebagai bahasa pemrograman.

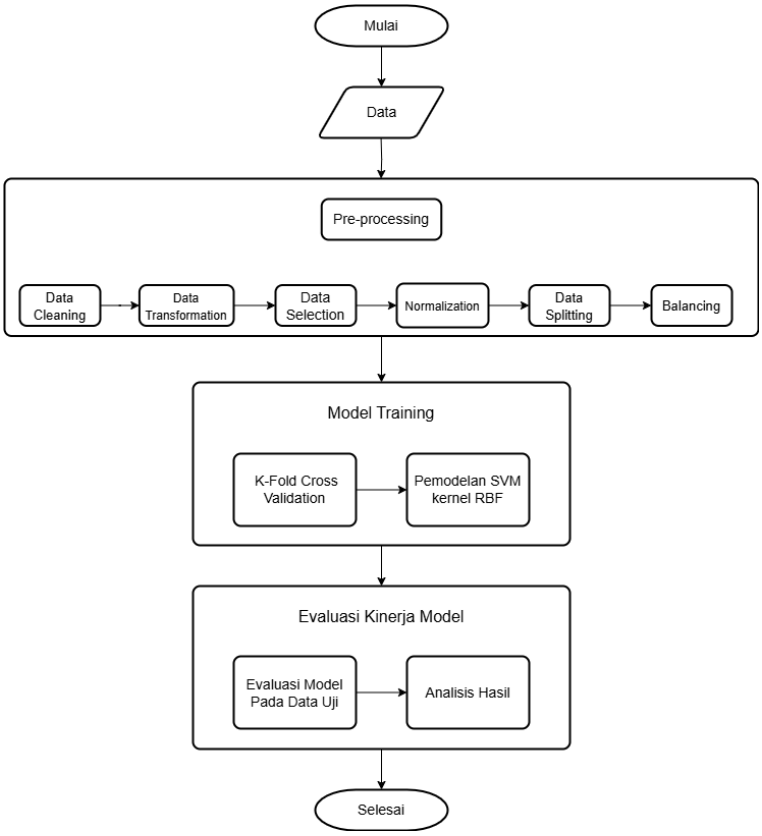
Data mentah didapatkan dari situs web kaggle.com (https://www.kaggle.com/datasets/nikhil1e9/loan-default/data) dan dilakukan proses seleksi atribut yang relevan. Terdapat 255.347 data dengan 18 atribut. Tabel 1 berikut ini mendeskripsikan dataset tersebut.

Tabel 1. Atribut dalam Dataset

No.	Atribut	Tipe	Keterangan
1	LoanID	String	Pengenal unik untuk setiap pinjaman.
2	Age	Integer	Umur peminjam.
3	Income	Integer	Pendapatan tahunan peminjam.
4	LoanAmount	Integer	Jumlah uang yang dipinjam.
5	CreditScore	Integer	Nilai kredit peminjam, yang menunjukkan kelayakan kredit mereka.
6	MonthsEmployed	Integer	Jumlah bulan peminjam telah bekerja.
7	NumCreditLines	Integer	Jumlah jalur kredit yang dimiliki peminjam.
8	InterestRate	Float	Tingkat suku bunga pinjaman.
9	LoanTerm	Integer	Jangka waktu pinjaman dalam bulan.
10	DTIRatio	Float	Rasio utang terhadap pendapatan, yang menunjukkan utang peminjam dibandingkan dengan pendapatan mereka.
11	Education	String	Tingkat pendidikan tertinggi yang dicapai oleh peminjam (PhD, Master's, Bachelor's, High school).
12	EmploymentType	String	Jenis status pekerjaan peminjam (Full-time, Part-time, Self-employed, Unemployed).
13	MaritalStatus	String	Status perkawinan peminjam (Single, Married, Divorced).
14	HasMortgage	String	Apakah peminjam memiliki hipotek (Yes dan No).
15	HasDependents	String	Apakah peminjam memiliki tanggungan (Yes dan No).
16	LoanPurpose	String	Tujuan pinjaman (Home, Auto, Education, Business, Other).
17	HasCoSigner	String	Apakah peminjam memiliki penandatangan bersama (Yes dan No).
18	Default	Integer	Variabel target biner yang menunjukkan apakah pinjaman tersebut gagal bayar (Yes dan No).

Gambar 1 merupakan gambaran umum penelitian yang dilakukan. Terdapat 3 tahapan besar yang dilakukan terhadap dataset yang menjadi input yaitu:

1. *Pre-processing* yang mencakup proses pembersihan data dari data yang hilang (*data cleaning*), transformasi data menjadi bentuk yang sesuai dengan algoritma SVM (*data transformation*), seleksi data untuk mendapat atribut yang relevan untuk diolah (d*ata selection*), normalisasi data agar nilai setiap atribut menjadi menjadi format yang lebih konsisten dan seimbang (*data normalization*), pemisahan data latih dan data uji (*data splitting*), serta penyeimbangan data (*data balancing*).
   Karena penelitian ini akan membandingkan pengaruh dari penyeimbangan data, maka dalam tahap *pre-processing* disiapkan 3 salinan dataset yaitu: a) dataset yang tidak dikenai penyeimbangan data, (b) dataset yang dikenai penyeimbangan data memakai ROS, (c) dataset yang dikenaik penyeimbangan data memakai NM.
2. *Model training* yang dilakukan dengan menerapkan model SVM dengan kernel RBF pada ketiga jenis dataset yang disiapkan pada tahap preprocessing. Validasi model dilakukan dengan menggunakan teknik k-fold cross validation dengan menggunakan tiga variasi nilai k yaitu 3, 5, dan 10 untuk dicari model terbaiknya.
3. *Evaluasi kinerja model:* model terbaik yang didapat dalam langkah 2 selanjutnya diterapkan pada data uji yang telah disisihkan sebelumnya.



Gambar 1. Gambaran umum penelitian

3. HASIL DAN PEMBAHASAN

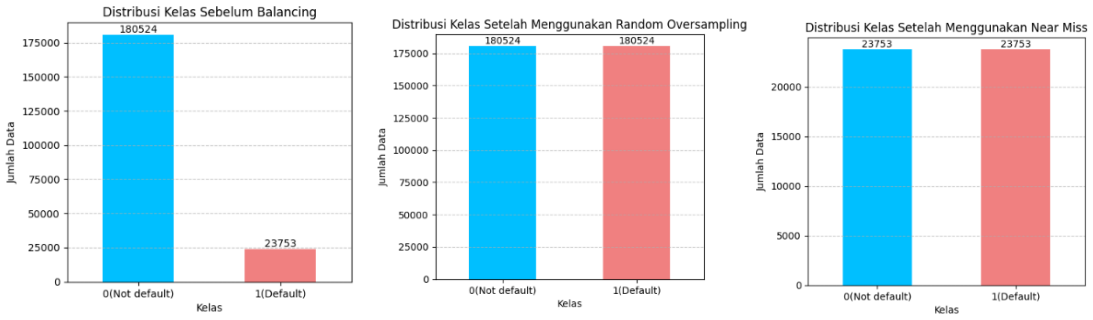
3.1. Tahap *Pre-processing*

Dari data mentah yang diinputkan, atribut LoanID dihapus karena tidak relevan untuk proses selanjutnya. Data dikenai proses *cleaning* terlebih dahulu dengan mengecek *missing values.* Dari hasil pengecekan tidak ditemukan adanya *missing values*. Selanjutnya dilakukan proses *data transformation* menggunakan *label encoder* yang berfungsi untuk mengubah nilai menjadi nilai numerik. Tahap *data selection* dilakukan untuk mendapatkan atribut yang paling relevan menggunakan *information gain.* Gambar 2 berikut ini menunjukkan hasil pemeringkatan *information gain*

	Information Gain
HasCoSigner	0.034853
HasDependents	0.034705
HasMortgage	0.033830
MaritalStatus	0.017839
Age	0.016047
Education	0.011610
EmploymentType	0.011598
NumCreditLines	0.011515
LoanTerm	0.010859
LoanPurpose	0.010412
InterestRate	0.008768
Income	0.007401
MonthsEmployed	0.005783
LoanAmount	0.003634
DTIRatio	0.000581
CreditScore	0.000487

Gambar 2. Hasil Pemeringkatan Atribut dengan Information Gain

Dari hasil pemeringkatan di atas, *DTIRation* dan *CreaditScore* dihapus karena nilai *information gain* dari 2 atribut ini memiliki perbandingan yang cukup jauh. Dengan demikian hanya 14 atribut yang digunakan dalam proses selanjutnya. Langkah *selanjutnya* adalah normalisasi menggunakan metode *min-max*. Normalisasi ini bertujuan untuk mengubah rentang nilai dari setiap atribut menjadi 0 hingga 1. Dengan normalisasi ini, perbedaan skala antar atribut dapat dihilangkan, sehingga memudahkan dalam proses analisis data lebih lanjut. Proses pemisahan data (*data splitting*) menjadi data latih dan data uji dilakukan dengan proporsi 80% data latih dan 20% data uji. Pada data latih dilakukan pengujian dampak penyeimbangan data terhadap akurasi model. Oleh karena itu maka disiapkan 3 jenis data latih yang berasal dari dataset seperti yang telah dijelaskan di bagian pendahuluan. Gambar 3 berikut ini adalah grafik perbandingan jumlah data kedua kelas sebelum dikenai proses penyeimbangan dan setelah dikenai proses penyeimbangan dengan ROS dan NM.



Gambar 3. Distribusi Kelas Sebelum dan Sesudah Proses Penyeimbangan Data

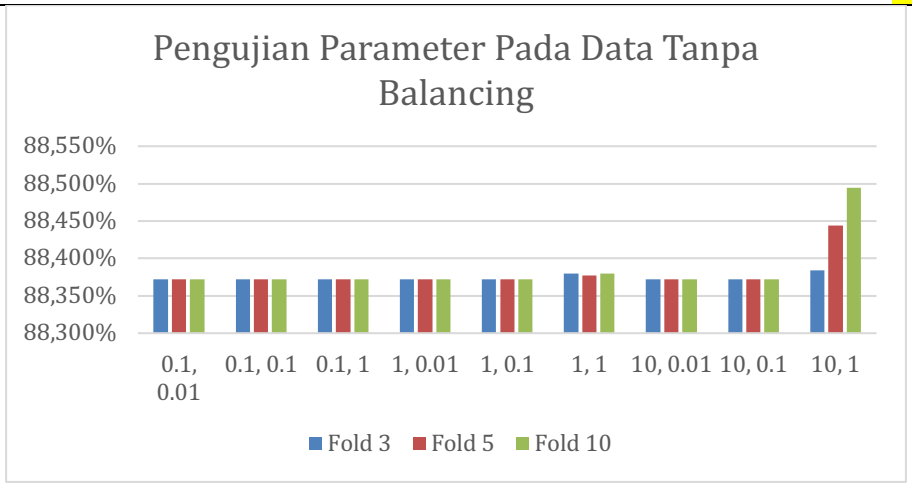
3.2. Tahap Pelatihan Model (*Model Training*)

Tahap selanjutnya adalah melakukan klasifikasi menggunakan algoritma *Support Vector Machine* (SVM) dengan kernel *Radial Basis Function* (RBF). Pengujian dilakukan dengan menggunakan beberapa variasi nilai C yaitu 0.1, 1, dan 10 serta beberapa variasi nilai gamma yaitu 0.01, 0.1, dan 1. Kinerja model diukur menggunakan teknik validasi silang k-fold dengan nilai K yang berbeda, yaitu 3, 5, dan 10. Indikator kinerja yang diukur adalah akurasi, presisi, *recall*, dan *f1-score.*

Tabel 1 berikut adalah tabel dan grafik hasil akurasi implementasi metode SVM pada 3 jenis dataset dengan 3 nilai fold.

Tabel 2. Hasil akurasi terhadap data tanpa penyeimbangan

Nilai K-Fold	C, Gamma								
	0.1, 0.01	0.1, 0.1	0.1, 1	1, 0.01	1, 0.1	1, 1	10, 0.01	10, 0.1	10, 1
3	88.37%	88.37%	88.37%	88.37%	88.37%	88.38%	88.37%	88.37%	88.38%
5	88.37%	88.37%	88.37%	88.37%	88.37%	88.38%	88.37%	88.37%	88.44%
10	88.37%	88.37%	88.37%	88.37%	88.37%	88.38%	88.37%	88.37%	88.49%

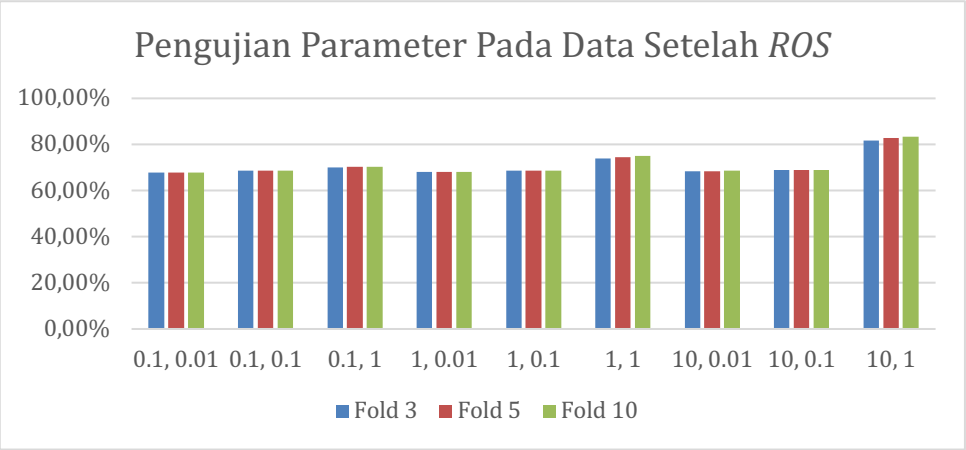


Gambar 4. Grafik hasil pengujian pada data tanpa penyeimbangan

P - ISSN: 1979-911X
E - ISSN: 2541-528X

Tabel 3. Hasil akurasi terhadap data yang diseimbangkan menggunakan Metode *Randon Over Sampling*

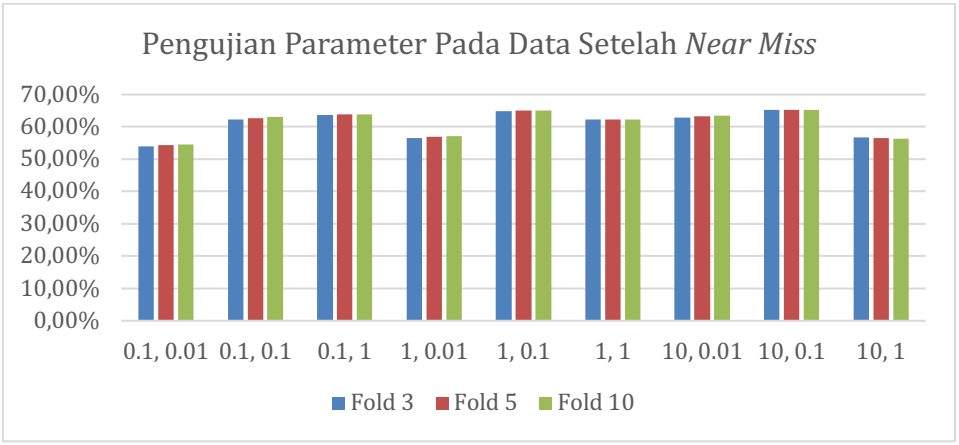
Nilai K-Fold	C, Gamma								
	0.1, 0.01	0.1, 0.1	0.1, 1	1, 0.01	1, 0.1	1, 1	10, 0.01	10, 0.1	10, 1
3	67.90%	68.55%	70.04%	68.11%	68.72%	73.92%	68.51%	69.00%	81.70%
5	67.92%	68.56%	70.22%	68.15%	68.74%	74.55%	68.52%	69.01%	82.83%
10	67.92%	68.56%	70.32%	68.19%	68.76%	74.94%	68.53%	69.01%	83.52%



Gambar 5. Grafik hasil pengujian pada data yang diseimbangkan dengan metode ROS

Tabel 4. Hasil akurasi terhadap data yang diseimbangkan menggunakan Metode *Near Miss*

Nilai K-Fold	C, Gamma								
	0.1, 0.01	0.1, 0.1	0.1, 1	1, 0.01	1, 0.1	1, 1	10, 0.01	10, 0.1	10, 1
3	54.00%	62.24%	63.59%	56.58%	64.93%	62.33%	62.85%	65.19%	56.69%
5	54.30%	62.69%	63.88%	56.90%	65.02%	62.34%	63.30%	65.20%	56.45%
10	54.49%	63.04%	63.92%	57.19%	65.08%	62.31%	63.47%	65.29%	56.32%



Gambar 6. Grafik hasil pengujian pada data yang diseimbangkan dengan metode NM

Pada dataset yang tidak dilakukan penyeimbangan didapatkan akurasi tertinggi sebesar 88.49% dengan menggunakan nilai C = 10, *gamma* = 1 pada *k-fold* = 10. Pada dataset yang dilakukan penyeimbangan menggunakan metode ROS didapatkan akurasi tertinggi sebesar 83.52% dengan menggunakan nilai C = 10, *gamma* = 1 pada *k-fold* = 10. Sedangkan pada dataset yang dilakukan penyeimbangan menggunakan NM didapatkan akurasi tertinggi sebesar 65.29% dengan menggunakan nilai C = 10, *gamma* = 0.1 pada k-fold = 10.

Tabel 5. Rangkuman hasil akurasi tertinggi beserta presisi dan recall

Data	C	Gamma	K-Fold	Akurasi	Presisi	Recall	F1-Score
Dataset tanpa penyeimbangan	10	1	10	88.49%	90%	10%	17%
Dataset setelah penyeimbangan menggunakan *ROS*	10	1	10	83.52%	84%	94%	89%
Dataset setelah penyeimbangan menggunakan *Near Miss*	10	0.1	10	65.29%	68%	62%	65%

Tabel 4 menunjukkan rangkuman hasil akurasi tertinggi yang dilakukan pada 3 jenis dataset disertai dengan nilai *precission, recall*, dan *F1-score*. Pada dataset tanpa penyeimbangan, akurasi tertinggi model SVM yang diperoleh dari parameter C=10, *gamma*=1, dan *k-fold*=10 adalah 88.49% dengan presisi 90%, namun nilai *recall* hanya sebesar 10% dan F1-score yang juga rendah yaitu 17%. Hal ini menunjukkan bahwa pada dataset tanpa penyeimbangan, model dapat mencapai akurasi yang tinggi hanya dengan memprediksi kelas mayoritas, karena sebagian besar sampel memang merupakan kelas mayoritas (Ghorbani & Ghousi, 2022). *Recall* yang rendah berarti banyak data positif yaitu kelas gagal bayar yang disebut sebagai kelas *Default* yang sebenarnya tidak terdeteksi. Nilai F1-score yang rendah juga menunjukkan keseimbangan antara presisi dan *recall* yang buruk.

Hasil pengujian terhadap dataset yang mengalami penyeimbangan dengan metode *Random Over Sampling* menghasilkan akurasi 83,52% yang diperoleh dengan menggunakan parameter C = 10 *gamma* = 1, dan *k-fold* = 10. Jika dibandingkan dengan dataset tanpa penyeimbangan, akurasi menurun sebesar 4.97%. Hal ini disebabkan oleh distribusi sampel dari dataset pelatihan yang mempengaruhi kinerja algoritma. Sebagai contoh, penambahan sampel duplikat untuk menyeimbangkan dataset menyebabkan SVM menjadi lebih bias terhadap beberapa sampel yang identik, yang mengakibatkan penurunan akurasi (Dina et al., 2022). Meskipun akurasinya lebih rendah dibandingkan dengan dataset yang tidak diseimbangkan, tetapi model dari dataset yang menggunakan ROS memiliki recall yang jauh lebih tinggi yaitu sebesar 94%. Ini menunjukkan kemampuan yang lebih baik dalam mendeteksi seluruh data gagal bayar (kelas *Default*) sebenarnya. Presisi sedikit menurun (84%) namun masih tetap tinggi menunjukkan model cukup akurat dalam prediksi positif. F1-score (89%) meningkat secara signifikan dibandingkan dataset tanpa penyeimbangan, menunjukkan keseimbangan yang jauh lebih baik antara presisi dan *recall*.

Pengujian terhadap dataset yang mengalami penyeimbangan dengan metode *Near Miss* menghasilkan akurasi terendah (65.29%) di antara ketiga dataset, namun *recall* masih cukup tinggi (62%), yang menunjukkan bahwa model masih mampu menangkap sebagian besar data positif sebenarnya meskipun akurasi menurun. Presisi sebesar 68%, menurun dibandingkan dua dataset sebelumnya, menunjukkan bahwa model menghasilkan lebih banyak kesalahan dalam prediksi positif. F1-score (65%) lebih rendah dibandingkan hasil penyeimbangan dengan ROS tetapi masih lebih tinggi daripada dataset tanpa penyeimbangan, menunjukkan peningkatan dalam keseimbangan presisi dan *recall*.

3.3. Tahap Evaluasi Kinerja Model

Pada pengujian menggunakan data *testing*, model SVM pada data tanpa penyeimbangan menunjukkan performa yang baik dengan akurasi 88.57%, presisi 56%, *recall* 5% dan *f1-score* 9%, sementara model pada data yang diseimbangkan dengan ROS menghasilkan akurasi sebesar 73.26%, presisi 21%, *recall* 47% dan *f1-score* 29%. Sedangkan penyeimbangan dengan *Near Miss* menghasilkan akurasi 43.50%, presisi 12%, *recall* 63% dan *f1-score* 21%. Pola kinerja model SVM pada data uji tampak serupa dengan pola kinerja model SVM pada data latih, di mana penyeimbangan dengan ROS meningkatkan recall tanpa terlalu mengorbankan presisi, menghasilkan F1-score terbaik dan menunjukkan performa yang seimbang dibandingkan dataset tanpa penyeimbangan maupun dataset yang diseimbangkan menggunakan NM.

4. KESIMPULAN

Berdasar hasil eksperimen menggunakan data latih, dataset tanpa penyeimbangan memberikan akurasi dan presisi tertinggi, tetapi recall yang sangat rendah menunjukkan bahwa model cenderung bias terhadap kelas mayoritas. Penyeimbangan dengan *Random Over Sampling* meningkatkan *recall* secara signifikan tanpa terlalu mengorbankan presisi, menghasilkan *F1-score* terbaik sebesar 89% dan menunjukkan performa yang seimbang. Penyeimbangan dengan *Near Miss* memberikan akurasi, presisi, dan *F1-score* yang lebih rendah dibandingkan ROS, namun tetap lebih baik dalam hal *recall* daripada dataset tanpa penyeimbangan. Eksperimen menggunakan

data uji menghasilkan pola kinerja yang serupa dengan data latih. Dapat disimpulkan bahwa penyeimbangan dengan ROS memberikan hasil terbaik dalam hal keseimbangan antara presisi dan *recall*, yang terlihat dari nilai F1-score tertinggi di antara ketiga metode.

DAFTAR PUSTAKA

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J., & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access*, *8*, 201173–201198.

Alamri, M., & Ykhlef, M. (2022). Survey of credit card anomaly and fraud detection using sampling techniques. *Electronics*, *11*(23), 4003.

Botchey, F. E., Qin, Z., & Hughes-Lartey, K. (2020). Mobile money fraud prediction—a cross-case analysis on the efficiency of support vector machines, gradient boosted decision trees, and naïve bayes algorithms. *Information*, *11*(8), 383.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

Dina, A. S., Siddique, A. B., & Manivannan, D. (2022). Effect of balancing data using synthetic data on the performance of machine learning classifiers for intrusion detection in computer networks. *IEEE Access*, *10*, 96731–96747.

Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access*, *8*, 67899–67911.

Hayder, I. M., Al Ali, G. A. N., & Younis, H. A. (2023). Predicting reaction based on customer's transaction using machine learning approaches. *International Journal of Electrical and Computer Engineering*, *13*(1), 1086.

Mallidi, M. K. R., & Zagabathuni, Y. (2021). Analysis of Credit Card Fraud detection using Machine Learning models on balanced and imbalanced datasets. *International Journal of Emerging Trends in Engineering Research*, *9*(7).

Najadat, H., Altiti, O., Aqoulehm, A. A., & Younes, M. (2020). Credit card fraud detection based on machine and deep learning. *11th International Conference on Information and Communication Systems (ICICS)*, 204–208.

Nalatissifa, H., Gata, W., Diantika, S., & Nisa, K. (2021). Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes, Support Vector Machine (SVM), dan Random Forest untuk Prediksi Ketidakhadiran di Tempat Kerja. *Jurnal Informatika Universitas Pamulang*, *5*(4), 578–584.

Pahlevi, O., Amrin, A., & Handrianto, Y. (2023). Implementasi Algoritma Klasifikasi Random Forest Untuk Penilaian Kelayakan Kredit. *Jurnal Infortech*, *5*(1), 71–76.

Sembiring, W. Y. M., Maulita, Y., & Ramadani, S. (2022). Pemamfaatan Metode Clustering Pada Nasabah Peminjaman Modal (Studi Kasus: PT. Faderal International Finance Binjai). *Jurnal Sistem Informasi Kaputama (JSIK)*, *6*(2), 346–356.

Singh, A., Ranjan, R. K., & Tiwari, A. (2022). Credit card fraud detection under extreme imbalanced data: a comparative study of data-level algorithms. *Journal of Experimental & Theoretical Artificial Intelligence*, *34*(4), 571–598.

Tamami, M. K., & Kharisudin, I. (2023). Komparasi Metode Support Vector Machine dan Naive Bayes Classifier untuk Pemodelan Kualitas Pengajuan Kredit. *Indonesian Journal of Mathematics and Natural Sciences*, *46*(1), 38–44.

Wibowo, P., & Fatichah, C. (2021). An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, *7*(1), 63–71.