Jurnal ICT : Information Communication & Technology
Vol. 20, N0.2, Desember 2021, pp. 365-372
p-ISSN: 2302-0261, e-ISSN: 2303-3363, DOI:

Penerapan Algoritma C4.5 Pada Imbalanced Dataset Untuk Memprediksi Kegagalan Angsuran Properti

Yodi Susanto[1], Muhammad Syafrullah[2], Devit Setiono[3]
[123] Fakultas Teknologi Informasi, Ilmu Komputer, Universitas Budi Luhur, Jakarta Selatan, Indonesia
Email: [1]yodi.susanto@gmail.com, [2]syafrullah@budiluhur.ac.id, devit.setiono@budiluhur.ac.id

INFORMASI ARTIKEL

Histori artikel:

Naskah masuk, 14 September 2021
Direvisi, 19 September 2021
Diiterima, 06 Oktober 2021

Kata Kunci:

Data Mining,
Prediction,
C.45,
K-NN,
Naïve Bayes

ABSTRAK

Abstract- *In this research, the data collection carried out by studying the patterns of consumers who fail to pay, which aimed to build a model so that it could be used in predicting customers who have the potential to fail to pay. The research used the Cross-Industry Standard Process for Data Mining (CRISP-DM) method with details of the business understanding process, data understanding, data preparation, modeling, evaluation and deployment / interpretation. The dataset in this research was taken from sales, cancellation and consumer data from January 2016 to December 2019. Because the dataset in this research was an imbalanced dataset, the researchers tried to use Synthetic Minority Oversampling Technique (SMOTE) in handling the imbalanced dataset. The research conducted a comparison of the value of accuracy, precision, recall, f measure and Area Under the ROC Curve (AUC) between the original dataset and the dataset for the addition of the SMOTE technique to several algorithms including C4.5, K-NN and Naïve Bayes. The attributes used in this research were source of funds, purpose of purchase, age, selling price, occupation, total installments, percentage of total installments, monthly installments, percentage of late installments and status. From the comparison, it was found that the C4.5 algorithm with the SMOTE 480% dataset had the highest accuracy value of 97.62%, precision of 0.976, recall of 0.976, f measure of 0.976 and AUC of 0.986 which meant Excellent Classification. From the research conducted, it was expected that the model formed on the imbalanced dataset with the C4.5 and SMOTE algorithms could be used to predict consumer installment failures.*

Abstrak- Pada penelitian ini akan dilakukan penambangan data dengan mempelajari pola dari konsumen yang gagal bayar, yang bertujuan untuk membangun model agar dapat digunakan dalam melakukan prediksi terhadapat kosumen yang berpotensi gagal bayar. Penelitian menggunakan metode *Cross-Industry Standard Process for Data Mining* (CRISP-DM) dengan perincian proses *business understanding, data understanding, data preparation, modeling*, *evaluation* dan *deployment/interpretation*. *Dataset* pada penelitian ini diambil dari data penjualan, pembatalan dan konsumen dengan periode Januari 2016 sampai dengan Desember 2019, dikarenakan *dataset* pada penelitian ini merupakan *imbalanced dataset* maka peneliti mencoba menggunakan *Synthetic Minority Oversampling Technique* (SMOTE) dalam menangani *imbalanced dataset* tersebut. Penelitian melakukan perbandingan nilai akurasi, presisi, *recall*, *f measure* dan *Area Under the* ROC *Curve* (AUC) antara *dataset* asli dengan *dataset* penambahan teknik SMOTE pada beberapa algoritma diantaranya C4.5, K-NN dan Naïve Bayes. Atribut yang digunakan pada penelitian ini adalah sumber dana, tujuan pembelian, umur, harga jual, pekerjaan, total angsuran, persentase jumlah angsuran bayar, angsuran perbulan, persentase jumlah terlambat angsuran dan status. Dari hasil perbandingan didapatkan bahwa algoritma C4.5 dengan *dataset* SMOTE 480% memiliki nilai akurasi tertinggi sebesar 97.62%, presisi 0.976, *recall* 0.976, *f measure* 0.976 dan AUC sebesar 0.986 yang berarti *Excellent Classification.* Dari penelitian

yang dilakukan ini diharapkan model yang terbentuk pada *imbalanced dataset* dengan algoritma C4.5 dan SMOTE dapat digunakan untuk melakukan prediksi gagal bayar angsuran konsumen.

***Penulis Korespondensi**:*

Muhammad Syafrullah
Program Studi Ilmu Komputer
Universitas Budi Luhur
Jl. Ciledug Raya, Petukangan Utara, Jakarta Selatan, 12260.
Email: syafrullah@budiluhur.ac.id

1. Pendahuluan

Bisnis properti di Indonesia saat ini sedang berkembang sangat pesat sekali, terutama pada kota-kota besar. Hal ini dikarenakan masih tingginya angka kebutuhan tempat tinggal yang belum terpenuhi atau *backlog*. Berdasarkan data Badan Pusat Statistik (BPS), *backlog* perumahan pada 2015 sebesar 11,4 juta unit atau menurun dari backlog pada 2010 yang mencapai 13,5 juta unit, data *backlog* 2010 tersebut berasal dari aspek kepemilikan rumah, terutama kebutuhan tempat tinggal bagi segmen menengah kebawah[1]. Atas hal tersebut banyak pengembang properti membangun hunian vertikal sebagai solusi untuk memenuhi kebutuhan tersebut dan juga dapat mengoptimalkan kebutuhan ruang terbuka hijau. Dalam proses penjualan unit properti, ada berbagai macam cara bayar yang dapat dipilih oleh konsumen seperti *Cash* Keras dan Kredit Kepemilikan Apartemen (KPA), beberapa pengembang besar menawarkan cara pembayaran Kontan Bertahap (KB) yaitu konsumen melakukan pembayaran angsuran kepada pihak pengembang langsung, hal ini sangat menarik minat bagi para konsumen, dikarenakan fasilitas angsuran dengan menggunakan skema cara bayar KB tidak memiliki bunga 0%.

Kemampuan konsumen dalam melakukan pembayaranpun harus dilakukan secara tepat waktu sesuai skema yang disetujui, agar tidak terjadi keterlambatan pembayaran yang mengakibatkan gagal bayar. Jika terjadi pembatalan unit, maka pihak pengembang akan dirugikan, karena akan mengakibatkan menurunnya proyeksi penerimaan serta sulitnya dalam menjual unit kembali karena harga unit sudah naik, hal ini pun akan berpengaruh terhadap pengelola unit karena pembebanan tagihan Iuran Pengelolaan Apartemen (IPL) akan ditanggung oleh pengelola (jika unit tersebut belum terjual). Namun selama periode pembayaran angsuran tersebut konsumen banyak yang tidak memenuhi kewajiban pembayarannya, yang mengakibatkan pembatalan unit properti yang dibeli. Dengan memanfaatkan *database* penjualan dari proyek properti yang berisi data penjualan konsumen, pembayaran dan administrasi, data tersebut dapat digali untuk mendapatkan pengetahuan dan pola dalam pengambilan keputusan dalam prediksi konsumen yang tidak menyelesaikan pembayaran angsuran dengan tepat waktu yang mengakibatkan pembatalan unit. *Data mining* adalah serangkaian proses untuk mendapatkan pengetahuan korelasi atau pola dari kumpulan data dengan memilih sejumlah besar data yang disimpan menggunakan teknologi pengenalan pola serta teknik statistik dan matematika [2].

Ada berbagai area pada *data mining* yang dapat dimanfaatkan dalam prediksi potensi gagal bayar angsuran, salah satunya melakukan analisis prilaku konsumen dalam pembayaran angsuran untuk menghindari gagal bayar dengan membandingkan algoritma C4.5, bayesNet dan Naïve Bayes [3], pada penelitian tersebut *dataset* yang digunakan memiliki tipe atribut nominal dan numerik, kemudian didapatkan bahwa algoritma C4.5 memiliki nilai akurasi tertinggi sebesar 78.3784%. Penelitian lainnya yang dilakukan oleh [4] menerapkan *data mining* untuk dapat membantu melakukan proses analisis kredit agar dapat menghasilkan informasi yang tepat, apakah nasabah yang akan mengajukan kreditnya memiliki status layak atau tidak layak, sehingga dapat melihat potensi pembayaran kredit yang dilakukan nasabah dengan menggunakan algoritma K-*Nearest Neighbor* (K-NN), dan didapatkan nilai akurasi tertinggi sebesar 93.33% dengan nilai k adalah 5. Berikutnya penelitian yang sudah dilakukan oleh [5] dengan memanfaatkan *data mining* menggunakan algoritma Naïve Bayes dan melakukan proses *feature selection* didapatkan nilai akurasi terbaik dalam model prediksi tingkat kelancaran pembayaran kredit sebesar 71.97%. Seperti yang terlihat pada gambar 1.3 bahwa *dataset* yang digunakan merupakan *imbalanced*

dataset, dimana terdapat data rasio yang tidak proposional pada kelas di dalam *dataset*, kondisi tersebut akan menimbulkan kemampuan model akan berkurang dalam mengenali pola saat proses analisis [6][7]. Untuk mengatasi *imbalanced dataset*, salah satu caranya menggunakan metode *over-sampling*, seperti pada kasus *Prediction of default payment of credit card clients using Data Mining Techniques* [8][9], dengan menerapkan teknik *over-sampling* yaitu *Synthetic Minority Oversampling Technique* (SMOTE) pada *imbalanced dataset* yang digunakan mampu meningkatkan kinerja model, sehingga mendapatkan nilai akurasi terbaik.

Berdasarkan penjelasan diatas, penelitian ini akan membuat model prediksi dengan melakukan komparasi algoritma C4.5, K-*Nearest Neighbor* (K-NN) dan Naïve Bayes serta menambahkan metode *over-sampling* yaitu teknik SMOTE untuk mengatasi *imbalanced dataset*, dimana model tersebut nantinya dapat digunakan oleh pihak manajemen melakukan analisis prediksi gagal bayar angsuran unit properti serta dapat mengambil keputusan untuk mencegah terjadinya gagal bayar, seperti melakukan *reschedule* skema pembayaran atau merubah cara bayar dengan menggunakan KPA.

2. Metode Penelitian

Penelitian kali ini menggunakan pendekatan eksperimen dengan cara menguji metode menggunakan data yang sama dengan bantuan tools WEKA. Dengan melakukan pengujian tersebut, akan didapatkan algoritma mana yang memiliki nilai akurasi tertinggi. Hasil pengujian nantinya dapat digunakan sebagai dasar evaluasi pihak manajemen kepada konsumen yang diprediksi akan gagal melakukan pembayaran kewajiban angsuran. Proses penambangan data dilakukan dengan menerapkan prosedur yang mengacu pada proses *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Merujuk pada proses CRISP-DM, penelitian ini akan menerapkan langkah-langkah penelitian sesuai dengan koridor tersebut.

1. Fase Pemahaman Bisnis (*Business Understanding)*

Proses ini dimulai dengan suatu pemahaman mengenai kebutuhan pihak pengelola terhadap *knowledge* baru dan suatu spesifikasi eksplisit dari tujuan bisnis mengenai penelitian ini. Kebutuhan mengenai bagaimana mendapatkan ciri-ciri umum dari konsumen yang sudah gagal bayar angsuran, menjadi alasan peneliti untuk mulai mempelajari data pola konsumen di proyek properti, melakukan pengumpulan data penjualan dan pembatalan unit, menganalisa dokumen pendukung dan mendapatkan keterangan dari pihak terkait.

2. Fase Pemahaman Data *(Data Understanding)*

Peneliti melakukan pengumpulan data yang berasal dari *database* penjualan dengan cara bayar Kontan Bertahap tahun 2016 sampai dengan 2019 sebanyak 2658 *record*, dengan 20 atribut prediktor yaitu tanggal beli, tahun beli, kode konsumen, sumber dana, tujuan pembelian, tanggal lahir, usia, kode unit, tipe unit, harga unit, cara bayar, tanggal batal, tahun batal, pekerjaan, jenis kelamin, angsuran terakhir, total angsuran, persentase angsuran, angsuran perbulan rupiah, jumlah terlambat. Beberapa hal yang dapat dilihat dari data tersebut salah satunya terkait data serapan pasar untuk pemakaian sendiri masih lebih besar jika dibanding dengan investor yang bertujuan membeli unit hanya untuk investasi.

3. Fase Pengolahan Data *(Data Preparation)*

Pada fase *Data Preparation* ini, peneliti melakukan preliminary research dan beberapa tahapan lainnya yang bertujuan untuk memperbaiki dan meningkatkan kualitas data, tahapan yang dilakukan oleh peneliti sebagai berikut:

a. Data *Cleansing*
   Data *cleansing* yang dimaksud adalah memperbaiki nilai data pada *missing value*, dimana nilai dari sebuah data pada atribut bernilai kosong atau tidak sesuai.
b. Data *Transformation*
   Transformasi data dilakukan dengan cara generalisasi data terhadap data yang sebenarnya memiliki kesamaan nilai atau arti.
c. Data *Reduction*
   Peneliti melakukan reduksi data berupa pengurangan jumlah data dan pengurangan atribut yang tidak relevan.
d. Pengujian Algoritma
   Dalam proses *preliminary research* yang dilakukan pada penelitian ini, peneliti melakukan pengujian algoritma dalam menangani dataset. Peneliti menggunakan algoritma C4.5, K-NN dan Naïve Bayes. Pemilihan algoritma tersebut merujuk pada beberapa penelitian yang sudah pernah dilakukan sebelumnya dengan mempertimbangkan kemiripan fokus penelitian.
e. Feature Selection

Feature Selection dilakukan dengan membandingkan beberapa metode dan penyesian parameter yang dianggap perlu, tujuan dari *Feature Selection* ini untuk mendapatkan atribut yang relevan sesuai dengan algoritma terpilih, agar mendapatkan model dengan nilai akurasi terbaik.

4. **Fase Pemodelan *(Modelling)***
   Pada fase modelling, beberapa teknik pemodelan dipilih dan diterapkan. Termasuk hasil dari preliminary research yang kemudian diolah lagi dengan menerapkan teknik SMOTE. Penambahan teknik SMOTE pada penelitian ini bertujuan untuk melihat apakah imbalanced dataset yang digunakan pada penelitian ini dengan teknik SMOTE, dapat meningkatkan nilai akurasi atau sebaliknya.

5. **Fase Evaluasi *(Evaluation)***
   Evaluasi dilakukan dengan confusion matrix pada satu atau lebih model yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektivitas. Evaluasi dilakukan terhadap nilai akurasi, presisi, recall, f-measure dan area under ROC. Tahap ini akan menetapkan apakah terdapat model yang dapat memenuhi tujuan penelitian. Selain itu, evaluasi dilakukan juga untuk menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik.

6. **Fase *Deployment / Interpretation***
   Fase ini menginterpretasikan hasil *data mining* dan model yang dihasilkan. Diharapkan model yang dihasilkan dapat berguna bagi perusahaan dalam kaitannya dengan proses bisnis yang dijalankan.

3. Hasil dan Pembahasan

3.1 Pengumpulan Data dan Analisi

Pengumpulan data pendukung selain dilakukan dengan observasi dan dokumentasi, peneliti juga melakukan wawancara dengan pakar. Pakar dalam hal ini adalah pihak manajemen proyek properti yaitu divisi marketing dan collection yang terkait langsung dengan proses penjualan. Dari wawancara tersebut, pihak pakar memberikan masukkan berupa beberapa atribut yang menurut pakar bisa menjadi faktor terjadinya gagal bayar.

3.2 *Data Collection*

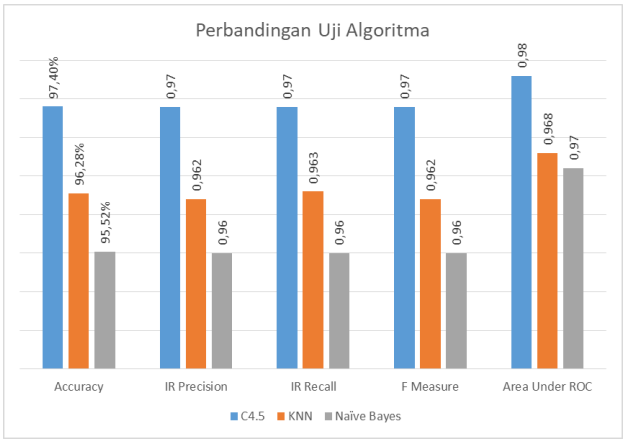
Pengumpulan *dataset* dilakukan dengan cara *query* pada *database* penjualan apartemen. *Query* digunakan peneliti untuk mempersempit pencarian data. Pencarian data difokuskan terhadap data yang relevan untuk mencari profil pembeli, master unit yang dijual serta riwayat pembayaran untuk mendapatkan pola gagal bayar penjualan unit di proyek properti.

3.3 Data *Preparation*

Dari data awal yang didapatkan sebanyak 2658 *record* dengan jumlah atribut sebanyak 20 atribut *predictor* dan 1 label kelas yaitu tanggal beli, tahun beli, kode *customer*, sumber dana, tujuan pembelian, tanggal lahir, usia, kode unit, tipe unit, harga unit, cara bayar, tanggal batal, tahun batal, pekerjaan, jenis kelamin, angsuran terakhir, total angsuran, persentase angsuran, angsuran perbulan rupiah, jumlah terlambat, peneliti melakukan beberapa proses *data preparation* yang bertujuan untuk meningkatkan kualitas data.

3.4 Uji Perbandingan Algoritma

Sebagai *preliminary research*, uji perbandingan algoritma dilakukan dengan membandingkan algoritma C4.5, K-NN dan Naïve Bayes dalam mengolah *dataset* yang digunakan pada penelitian ini. Pengujian dilakukan dengan menggunakan mode 10 *Fold cross-validation* dengan keluaran berupa nilai akurasi, *precision, recall, F-Measure dan* AUC. Untuk algoritma KNN, peneliti menggunakan nilai k=1, hal ini dipilih karena peneliti sudah melakukan perbandingan nilai k (1,3,5,7,9) pada algortima KNN dan nilai k=1 merupakan nilai akurasi tertinggi pada algortima KNN. Hasil perbandingan dapat dilihat pada gambar 1.



Gambar 1. Perbandingan Uji Algoritma

3.5 *Feature Selection*

Kemudian pada tahapan ini juga dilakukan *feature selection* dengan menggunakan *wrapper method* dan *ranker*. *Feature selection* dilakukan dengan beberapa *atribut evaluator* dan metode. Hasil dari perbandingan *Feature Selection* dapat terlihat pada gambar 2.

	A	B	C	D	E	F
Search Method	Best First	Best First	Best First	Best First	Attribute Ranking	Attribute Ranking
Search Direction	Backward	Forward	Backward	Forward		
Atribut Evaluator	Wrapper Subset	Wrapper Subset	Wrapper Subset	Wrapper Subset	Gain Ratio Feature	Info Gain Ratio Feature
Selection Mode	Full Training Set	Full Training Set	10 fold CV	10 fold CV		
Selected Atributes	tujuanbeli	tujuanbeli	sumberdana (0%)	sumberdana (0%)	angs_prs (0.210133)	angs_prs (0.31587)
	tipeunit	tipeunit	tujuanbeli (100%)	tujuanbeli (100%)	tujuanbeli (0.142303)	tujuanbeli (0.134547)
	hargaunit	hargaunit	usia (40%)	usia (50%)	prs_terlambat_lbh_30 (0.018709)	prs_terlambat_lbh_30 (0.008026)
	pekerjaan	pekerjaan	tipeunit (100%)	tipeunit (90%)	hargaunit (0.005822)	hargaunit (0.006616)
	jk	jk	hargaunit (20%)	hargaunit (20%)	totalang (0.003767)	totalang (0.005793)
	totalang	totalang	pekerjaan (100%)	pekerjaan (100%)	pekerjaan (0.002015)	pekerjaan (0.004232)
	angs_prs	angs_prs	jk (30%)	jk (30%)	angperbulan_rp (0.001803)	usia (0.003414)
	angperbulan_rp	angperbulan_rp	totalang (100%)	totalang (90%)	usia (0.001769)	angperbulan_rp (0.002596)
	prs_terlambat_lbh_30	prs_terlambat_lbh_30	angs_prs (100%)	angs_prs (100%)	sumberdana (0.000688)	sumberdana (0.000981)
			angperbulan_rp (100%)	angperbulan_rp (100%)	jk (0.000521)	jk (0.000521)
			prs_terlambat_lbh_30 (100%)	prs_terlambat_lbh_30 (100%)	tipeunit (0.000332)	tipeunit (0.000353)

Gambar 2. Perbandingan *Feature Selection*

Peneliti kemudian melakukan simulasi menggunakan algoritma C4.5 terhadap hasil *feature selection* tersebut untuk mendapatkan nilai akurasi terbaik. Simulasi dilakukan dengan mereduksi pada kolom C, D, E dan F. Untuk atribut pada kolom C dan D pada tabel 4.13, peneliti melakukan uji coba dengan mereduksi atribut secara bertahap, mulai dari menguji atribut dengan nilai diatas 10%, diatas 20%, diatas 30%, diatas 40%, diatas 50% dan diatas 90%. Pada hasil uji coba yang dilakukan tersebut, peneliti mendapatkan bahwa atribut diatas 30% memiliki nilai tertinggi, sehingga pada kolom C dan D, atribut "sumberdana", "hargaunit" dan "jk" direduksi karena memiliki nilai dibawah sama dengan 30%. Pada kolom E, peneliti melakukan uji coba dengan atribut diatas 0.000332 sampai dengan diatas 0.142303, dari hasil uji tersebut didapatkan bahwa atribut diatas 0.000521 memiliki nilai akurasi terbesar, sehingga pada kolom E dilakukan reduksi terhadapt dua atribut yaitu "jk" dan "tipeunit". Pada kolom F peneliti melakukan uji coba dengan cara mereduksi atribut yang dimulai dengan nilai diatas 0.000353 sampai dengan nilai atribut diatas 0.134547, hasil pada uji coba kolom F, didapatkan atribut dengan nilai diatas 0.000521, sehingga pada kolom F dilakukan reduksi sebanyak 2 atribut, yaitu "jk" dan "tipeunit".

Setelah dilakukan reduksi atribut, peneliti melakukan pengujian terhadap atribut terpilih pada tabel 4.14 dengan menggunakan algoritma C4.5 dengan *Cross Validation* 10 *fold*. Terlihat pada tabel 4.15 setelah dilakukan pengujian, didapat pada kolom E dan F memiliki nilai akurasi tertinggi sebesar 97.52%, dimana atribut pada kolom E dan F memiliki atribut yang sama.

Tabel 1. Hasil Uji Atribut Terpilih

Comparison Field	*A*	*B*	*C*	*D*	*E*	*F*
Accuracy	97.48%	97.48%	97.48%	97.48%	97.52%	97.52%
IR Precision	0.975	0.975	0.975	0.975	0.975	0.975
IR Recall	0.975	0.975	0.975	0.975	0.975	0.975
F Measure	0.975	0.975	0.975	0.975	0.975	0.975
Area under ROC	0.983	0.983	0.982	0.982	0.978	0.978

3.6 Modelling

Pada penelitian ini, peneliti melakukan tahapan *modelling* yang diawali dengan melakukan *prelimenary research* menggunakan aplikasi WEKA. Hasil dari *prelimenary research* yang sudah dilakukan didapatkan algoritma C4.5 sebagai algoritma terpilih dengan 9 atribut *predictor* dan 1 label *class*, dimana atribut tersebut merupakan hasil tahapan pada proses *feature selection*. Model yang dibuat dengan algoritma C4.5 akan diuji dengan menerapkan teknik SMOTE untuk mengatasi *imbalanced dataset*.

3.7 Penambahan SMOTE Dataset

Pada tahapan ini, peneliti melakukan uji coba pembuatan model dengan menggunakan *dataset* yang telah dilakukan proses *feature selection*, dimana terdapat 9 atribut *predictor* dan 1 label *class* didalamnya. Peneliti akan menggunakan teknik *over-sampling* dengan SMOTE, yaitu menambahkan nilai persentase pada kelas minor, agar kelar minor dapat seimbang dengan kelas mayor. Untuk mencari nilai persentase kenaikan kelas minor, penelitian sebelumnya melakukan penambahan persentase pada data minor dengan patokan banyaknya data mayor [10], pada peneliti ini, peneliti melakukan beberapa kali percobaan dengan cara menaikan nilai persentase kelas minor hingga mendapatkan hasil maksimal dari nilai akurasi yang didapat dan banyaknya data minor sama dengan data mayor. Perbandingan hasil uji coba yang dilakukan pada *dataset* setelah *feature selection* dengan penambahan teknik SMOTE.

Dari hasil pengujian yang sudah dilakukan, nilai akurasi tertinggi dengan menggunakan SMOTE berada pada 97.62% dengan penambahan data minor sebesar 480%, dimana data minor menjadi sebanyak 2261 *record*. Jika dibandingkan dengan data tanpa SMOTE, terdapat peningkatan nilai akurasi sebesar 0.10%.

3.8 Model Tanpa Menerapkan Data SMOTE

Pada model tanpa SMOTE, total terdapat 2658 *record*, dari total data tersebut terdapat 390 *record* yang merupakan kelas minor (label *class* "BATAL") dan 2268 *record* yang merupakan kelas mayor (label *class* "LANCAR"). Pengukuran dilakukan pada *dataset* yang diolah dengan menggunakan algoritma C4.5. Atribut yang digunakan adalah atribut awal dengan 9 *predictor* dan 1 label *class*. Pengukuran dilakukan pada algoritma C4.5 dengan mode *Cross Validation* 10 *fold*, sehingga menghasilkan *confusion matrix* yang terdiri dari nilai *accuracy*, *precision dan recall* dapat dilihat pada tabel 2.

Tabel 2. Hasil Explorer C4.5 Tanpa SMOTE

Explore	*Accuracy*	*IR Precision*	*IR Recall*	*Area under ROC*	*F Measure*	*Class*
C4.5 10 Cross-validation 9 Atribut predictor 1 Label class	97.5169%	0.985	0.986	0.978	0.985	LANCAR
		0.918	0.913	0.978	0.915	BATAL

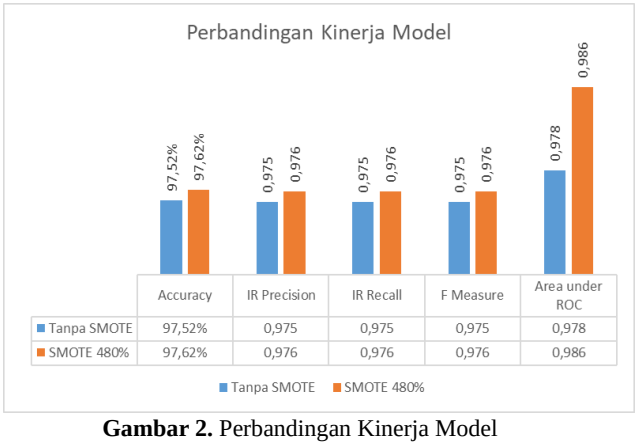
3.9 Model Setelah Menerapkan SMOTE

Pada model dengan SMOTE, terlihat pada tabel 4.16, dilakukan peningkatan data minor sebesar 480% dengan teknik SMOTE yang menghasilkan data minor sebanyak 2261 *record*,sehingga data minor mendekati seimbang dengan data mayor. Pengukuran *confusion matrix* kemudian dilakukan terhadap *dataset* yang sudah melalui teknik SMOTE. Dari tabel 4.16 mengenai hasil perbandingan penerapan SMOTE pada *imbalaced dataset*, didapatkan nilai akurasi tertinggi sebesar 97,62%. Nilai tersebut didapatkan dengan menerapkan SMOTE sebesar 480% yang diuji dengan algoritma C4.5. Hasil pergukuran mandiri tersebut dapat dilihat pada tabel 3.

Explore	*Accuracy*	*IR Precision*	*IR Recall*	*Area under ROC*	*F Measure*	*Class*
C4.5 10 Cross-validation 9 Atribut predictor 1 Label class	97.6154%	0.976	0.977	0.986	0.976	LANCAR
		0.977	0.976	0.986	0.976	BATAL

Tabel 3. Hasil Explorer C4.5 Dengan SMOTE

3.10 Perbandingan Kinerja Model



Gambar 2. Perbandingan Kinerja Model

3.11 Interpretasi

Setelah dilakukan proses pemodelan dan didapatkan hasil nilai akurasi sebesar 97.62% dengan penambahan teknik SMOTE sebesar 480% pada algoritma C4.5. Dengan nilai akurasi tersebut jika model mampu melakukan prediksi gagal bayar konsumen secara tepat, maka model tersebut dapat membantu pihak manajemen proyek properti dalam melakukan strategi untuk mencegah terjadinya pembatalan unit dari konsumen, seperti melakukan negosiasi penjadwalan ulang waktu pembayaran yang lebih panjang, perpindahan cara bayar menjadi KPA ataupun pemindahan unit dengan harga yang lebih rendah. Namun sebaliknya jika prediksi yang dilakukan oleh model meleset, seperti konsumen yang diprediksi lancar namun kenyataannya konsumen tersebut gagal bayar dan batal, maka potensi kerugian yang dialami oleh manajemen pun akan tetap tidak terhindarkan.

4. Kesimpulan

Setelah dilakukan uji coba dengan penambahan teknik SMOTE pada imbalanced dataset, dimana hasil uji coba dengan menambahkan data minor sebesar 480%, sehingga data minor mendekati seimbang dengan data mayor, dengan data minor awal sebanyak 390 menjadi 2261 dan data mayor 2268 menghasilkan nilai akurasi sebesar 97.62%, sehingga terbukti dengan penambahan teknik SMOTE yaitu menambahkan data minor sebesar 480% dapat meningkatkan nilai akurasi sebesar 0.10% dari nilai akurasi sebelumnya sebesar 97.52% pada algoritma C4.5. Kemudian untuk nilai presisi, recall, F measure dan AUC terjadi kenaikan setelah dilakukan penambahan teknik SMOTE, dimana untuk nilai presisi, recall dan F measure mengalami peningkatan sebesar 0.001, sedangkan untuk nilai AUC mengalami peningkatan sebesar 0.008. Hasil nilai presisi, recall dan F measure sebesar 0.976, sedangkan hasil nilai AUC sebesar 0.986, dimana nilai AUC tersebut dapat dikategorikan sebagai Excellent Classification. Prototipe dibuat berdasarkan model prediksi yang terbentuk. Prototipe tersebut memiliki kemampuan melakukan prediksi klasifikasi dengan kelas "Gagal" maupun "Lancar". Prediksi dilakukan dengan memberikan 11 nilai atribut yaitu Kode Unit, Nama Konsumen, Sumber Dana, Tujuan Pembelian, Usia, Harga Unit, Pekerjaan, Angsuran Terakhir, Total Angsuran, Jumlah Angsuran per Bulan dan Jumlah Angsuran Terlambat. Dengan kemampuan prototipe yang telah dikembangkan, yaitu untuk melakukan prediksi gagal bayar, diharapkan dapat membantu manajemen properti dalam membuat perencanaan dan strategi untuk mengatasi potensi gagal bayar yang mungkin terjadi dan perusahaan dapat terhindar dari resiko kerugian. Beberapa hal untuk penelitian selanjutnya yang harus diperhatikan adalah ekplorasi model prediksi dapat ditambahkan atribut lainnya dari profil konsumen seperti jumlah penghasilan, hal ini untuk mengetahui tingkat kemampuan konsumen dalam membeli dan mencicil pembayaran, pemodelan dapat menggunakan algoritma dan penambahan teknik lainnya untuk menangani *imbalanced dataset*, seperti melalukan uji coba dengan metode *ensemble* dan proses *feature selection* dapat menggunakan metode lainnya, sehingga didapatkan atribut dengan nilai akurasi terbaik.

Daftar Pustaka

[1] M. Roestamy and R. Rahmawati, "Model Pengembangan Paradigma Masyarakat bagi Kepemilikan Rumah yang Terpisah dari Tanah," *Mimb. Huk. - Fak. Huk. Univ. Gadjah Mada*, vol. 30, no. 2, p. 331, 2018, doi: 10.22146/jmh.17646.

[2] D. T. Larose and C. D. Larose, *DISCOVERING KNOWLEDGE IN DATA An Introduction to Data Mining Second Edition Wiley Series on Methods and Applications in Data Mining*. 2014.

[3] A. Jafar Hamid and T. M. Ahmed, "Developing Prediction Model of Loan Risk in Banks Using Data Mining," *Mach. Learn. Appl. An Int. J.*, vol. 3, no. 1, pp. 1–9, 2016, doi: 10.5121/mlaij.2016.3101.

[4] T. T. Muryono and I. Irwansyah, "Implementasi Data Mining Untuk Menentukan Kelayakan Pemberian Kredit Dengan Menggunakan Algoritma K-Nearest Neighbors (K-NN)," *Infotech J. Technol. Inf.*, vol. 6, no. 1, pp. 43–48, Jun. 2020, doi: 10.37365/it.v6i1.78.

[5] M. Hasan, "Prediksi Tingkat Kelancaran Pembayaran Kredit Bank Menggunakan Algoritma Naïve Bayes Berbasis forward Selection," vol. 9, pp. 317–324, 2017.

[6] D. M. Tampubolon, "Evaluasi Performa Kredit Menggunakan Data Mining untuk menilai permohonan Kredit Fasilitas Layanan Pembiayaan Perumahan: Studi Kasus PT. Bank XYZ," 2017.

[7] Novakovic, J., Veljovi, A., Iiic, S., Papic, Z. dan Tomovic, M. (2017) "Evaluation of Classification Models in Machine Learning," *Theory and Applications of Mathematics & Computer Science*, 7(1), hal. 39–46.

[8] A. Subasi and S. Cankurt, "Prediction of default payment of credit card clients using Data Mining Techniques," *Proc. 5th Int. Eng. Conf. IEC*

2019, pp. 115–120, 2019, doi: 10.1109/IEC47844.2019.8950597.

[9] Taufiq, I., Nur, A., Setiawan, N. Y. dan Bachtiar, F. A. (2018) "Prediksi Kredit Macet Berdasarkan Preferensi Nasabah Menggunakan Metode Klasifikasi C4 . 5 pada Koperasi Simpan Pinjam Mitra Raya Wates," J-ptiik, 2(12), hal. 6118–6127.

[10] Kasanah, A. N., Muladi, M. dan Pujianto, U. (2019) "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 3(2), hal. 196–201. doi: 10.29207/resti.v3i2.945.