EULER: Jurnal Ilmiah Matematika, Sains dan Teknologi
Journal Homepage: http://ejurnal.ung.ac.id/index.php/Euler
DOI: https://doi.org/10.34312/euler.v10i2.16751
Desember 2022, Vol. 10, No. 2, Hal. 262-273

Perbandingan Metode *k-Nearest Neighbor*, Regresi Logistik Biner, dan Pohon Klasifikasi pada Analisis Kelayakan Pemberian Kredit

Shantika Martha[1], Wirda Andani[2]*, Setyo Wira Rizki[3]

[1,2,3]Program Studi Statistika, Jurusan Matematika, Universitas Tanjungpura, Pontianak 78124, Indonesia
**Penulis Korespondensi. Email: wirda.andani@math.untan.ac.id*

Abstrak

Kredit Tanpa Agunan (KTA) adalah salah satu jenis kredit yang dapat dimanfaatkan oleh debitur tanpa jaminan. Beberapa debitur yang sudah melakukan KTA namun masih memerlukan tambahan pinjaman dana dapat melakukan *top up*. Namun pemberian KTA *top up* tidak terlepas dari risiko kredit macet atau risiko gagalnya debitur dan/atau pihak lain dalam menuntaskan hutangnya kepada bank. Dalam upaya penilaian kelayakan calon debitur maka bank membutuhkan suatu cara atau alat yang efektif untuk memprediksi manakah debitur yang mampu menyelesaikan pembayaran kredit dengan lancar. Alat bantu yang sering digunakan adalah metode klasifikasi. Pada penelitian ini dilakukan perbandingan tiga metode klasifikasi yaitu *k-nearest neighbor*, regresi logistik biner dan pohon klasifikasi untuk mendapatkan metode terbaik dalam menganalisis kelayakan pemberian KTA *top up*. Berdasarkan nilai *accuracy* pada masing-masing metode, metode pohon klasifikasi menghasilkan nilai *accuracy* yang paling tinggi dibandingkan kedua metode lainnya. Dengan demikian, untuk penelitian ini metode pohon klasifikasi merupakan metode yang paling baik dengan nilai akurasi sebesar 87,68%. Berdasarkan metode pohon klasifikasi, peubah peubah yang digunakan dalam aturan di pohon klasifikasi adalah DBR, lama bekerja seorang debitur, limit kredit, jenis pekerjaan debitur, jumlah penghasilan debitur, wilayah tempat tinggal debitur serta jangka waktu kredit debitur selama 1 bulan.

Kata Kunci: Kredit Tanpa Angunan; Metode Klasifikasi; *k-Nearest Neighbor*; Regresi Logistik Biner; Pohon Klasifikasi

Abstract

Kredit Tanpa Anggunan (KTA) are bank loans given to debtors without asking for guarantees. Some debtors who have made KTA but still need additional loan funds can top up. However, offering this facility to the public cannot be separated from the risk that the debtor and/or other parties fail to fulfill their obligations to the bank. In an effort to assess the feasibility of prospective debtors, banks need decision-making tools so that they can easily and quickly estimate which debtors are able to pay off credit on time (good credit). The tool that is often used is classification. In this study, we will compare 3 classification methods, namely k-nearest neighbor, binary logistic regression, and classification tree, to obtain the best method for analyzing the feasibility of giving KTA top-up. Based on the accuracy value in each method, the classification tree produces the highest accuracy value compared to the other two methods. Thus, for this study, the classification tree is the best method, with an accuracy value of 87.68%. The variables used in the classification tree are DBR, length of service of a debtor, credit limit, type of debtor's occupation, the total income of the debtor, the area where the debtor lives, and the credit period of the debtor is 1 month.

Keywords: *Kredit Tanpa Angunan; Classification; k-Nearest Neighbor; Binary Logistic Regression; Classification Tree*

1. Pendahuluan

Saat ini bank adalah salah satu lembaga keuangan yang esensial untuk memenuhi kebutuhan hidup masyarakat. Bank menyuguhkan suatu fasilitas dimana masyarakat ataupun badan usaha dapat meminjam sejumlah uang dan peminjam diwajibkan untuk melunasi hutangnya dengan mengangsur

Submit: 15 Oktober 2022, Disetujui: 21 November 2022, Online:25 November 2022

dengan pemberian bunga dalam waktu yang ditentukan [1]. Fasilitas ini sangat di minati masyarakat maupun perusahaan.

Layanan jasa kredit adalah salah satu fasilitas yang disediakan oleh bank dengan meminjamkan uang kepada seseorang maupun badan usaha dan mewajibkan pihak peminjam untuk melunasi utangnya secara mengangsur dengan pemberian bunga dalam jangka waktu yang ditentukan [1]. Ketersediaan dari layanan kredit ini sangat dibutuhkan karena jika kita perhatikan, di zaman sekarang kebutuhan semua orang selalu meningkat setiap tahunnya. Sebelum dapat menggunakan layanan ini, bank akan meminta jaminan dari seorang debitur. Namun jika debitur tidak mempunyai jaminan, bank memberikan solusi yaitu kredit tanpa anggunan. Kredit Tanpa Agunan (KTA) adalah fasilitas kredit yang diberikan oleh bank tanpa meminta agunan atau jaminan.

Ternyata terdapat beberapa debitur yang sudah melakukan KTA namun masih memerlukan tambahan pinjaman dana. Solusi permasalahan ini juga disediakan oleh pihak bank, yang disebut *top up* yaitu fasilitas khusus yang dapat digunakan debitur untuk mengajukan atau menambah pinjaman saat kreditnya belum lunas. Namun pemberian KTA *top up* tidak terlepas dari risiko kredit macet atau risiko gagalnya debitur dan/atau pihak lain dalam menuntaskan hutangnya kepada bank. Maka dari itu, penerimaan calon nasabah kredit harus selektif untuk meminimalisir terjadinya keterlambatan debitur membayar hutangnya (kredit macet) [2]. Dalam upaya penilaian kelayakan calon debitur maka bank membutuhkan suatu cara atau alat yang efektif untuk memprediksi manakah debitur yang mampu menyelesaikan pembayaran kredit dengan lancar. Alat bantu yang sering digunakan adalah metode klasifikasi.

Metode klasifikasi memungkinkan peneliti untuk mengklasifikasikan suatu objek baru ke dalam kelas tertentu berdasarkan nilai atribut-atributnya [3]. Algoritma pada metode klasifikasi terbagi menjadi 2 yaitu *Unsupervised learning* dan *Supervised learning.* Algoritma *Unsupervised learning* adalah suatu proses penemuan model (atau fungsi) yang membedakan kelas data sehingga bertujuan untuk digunakan memprediksi kelas dari objek yang label kelasnya tidak diketahui [4]. Sedangkan algoritma *Supervised learning* adalah algoritma untuk memperoleh "aturan" bagi penentuan keanggotaan kelas dari amatan lainnya nanti. Dalam rangka memprediksi calon debitur yang mengajukan kredit dapat menggunakan metode klasifikasi dengan algoritma *supervised learning* sehingga meminimalisir terjadinya kredit macet setelah pengajuan kredit disetujui.

Penelitian sebelumnya oleh Y. Maldini, A. M siregar dan T. A Mudzakir [5] melakukan penelitian analisis kelayakan pemberian kredit bagi nasabah koperasi menggunakan algoritma C4.5 dan KNN. Dari penelitian tersebut didapatkan kesimpulan bahwa KNN lebih baik dalam menentukan pemberian kredit kepada nasabah koperasi dibandingkan Algoritma C4.5. S. Sreesouthry, A. Ayubkhan , M. M. Rizwan , D. Lokesh dan K. P. Raj [6] juga melakukan penelitian untuk memprediksi calon debitur yang mengajukan kredit menggunakan Regresi Logistik dengan tingkat akurasi 77%. Dilakukan pula penelitian yang sejenis oleh M. Madaan, A. Kumar, C. Keshri , R. Jain dan P. Nagrath [7] dengan membuat perbandingan metode yaitu Random Forest dan Pohon klasifikasi. Pada penelitian ini penulis tertarik untuk melakukan analisis klasifikasi debitur KTA (kredit tanpa agunan) berpotensi melakukan *top-up* kredit atau tidak dengan risiko salah pengklasifikasian seminimal mungkin dengan membandingkan beberapa metode yaitu KNN, Regresi logistik dan Pohon klasifikasi. Dari ketiga metode ini ditentukan metode terbaik berdasarkan nilai akurasi model. Selain itu, ingin diketahui pula peubah penting yang mempengaruhi kelayakan pemberian kredit.

2. Metode Penelitian

Data yang digunakan adalah data sekunder, bersumber dari salah satu Bank di Indonesia dengan total jumlah debitur sebanyak 94000 debitur. Pada data tidak ditemukan data hilang atau adanya ketidak konsistenan data sehingga tidak dilakukan *cleaning data*. Adapun rincian peubah yang akan di gunakan dijelaskan pada Tabel 1.

Tabel 1. Detail Peubah yang digunakan

Peubah	Nama Peubah	Keterangan Konten Peubah
Y	Status Upselling	Top Dan New
X1	Jenis Kelamin	Laki-Laki dan Perempuan
X2	Usia	-
X3	Status Pernikahan	Belum Menikah, Menikah dan Cerai
X4	Jumlah Tanggungan	-
X5	Status Tempat Tinggal	Rumah Milik Sendiri, Rumah Milik Orang Tua, Rumah Dinas, Rumah Sewa, dan Rumah Kredit
X6	Pendidikan	SD, SMP, SMA, Diploma, S1, S2/S3
X7	Wilayah	Medan, Palembang, Bandung, Jakarta, Semarang, Surabaya, Denpasar, Balikpapan, Makassar Jayapura,
X8	Jenis Pekerjaan	Pegawai Swasta,Pegawai Bumn, Pns
X9	Lama Bekerja	-
X10	Penghasilan	-
X11	DBR (*Debt Burden Ratio*)	-
X12	Jangka Waktu Kredit Dalam Bulan	-
X13	Limit Kredit	-

Proses analisis data menggunakan *software R Studio*, dimulai dengan melakukan deskripsi data untuk melihat gambaran umum data. Setelah itu dalam menyusun model tersebut, 94.000 data debitur dibagi secara acak menjadi dua gugus data yaitu 80000 data latih (data *training*) dan sisanya 14000 data uji (*testing*). Langkah terakhir yaitu membangun tiga (3) model klasifikasi yaitu *k-nearest neighbor*, regresi logistic biner dan pohon klasifikasi.

2.1 K-Nearest Neighbor

K-Nearest Neighbor adalah salah satu metode klasifikasi yang mengategorikan data baru berdasarkan kedekatan lokasi (jarak) paling dekat suatu data baru dengan data lain atau kebeberapa data/tetangga (*neighbor*) terdekat [8]. Terdapat beberapa cara untuk mengukur jarak kedekatan antara data baru dengan data lama, untuk penelitian ini penulis menggunakan *euclidean distance* [4]:

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

Dalam penggunaan metode ini harus berhati hati pada satuan data yang ingin dianalisis. Biasanya data yang digunakan mempunyai variabilitas satuan sehingga mengakibatkan hasil analisis menjadi tidak tepat. Maka sebelum melakukan pengklasifikasian ada baiknya melakukan standardisasi atau tansformasi data. Pada penelitian ini penulis menggunakan *min-max normalization* karena data terdiri dari skala pengukuran numerik dan kategorik. Jika data *training* terdiri dari campuran skala pengukuran antara numerik dan kategori, lebih baik gunakan *min-max normalization* [9]:

$$x_{baru} = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

dimana

$i = 1,2, \dots, p$, p = banyak peubah penjelas.

Setelah data di transformasi menjadi satuan yang sama, selanjutnya proses yang tak kalah penting karena dapat mempengaruhi tingkat akurasi dari model untuk memprediksi yaitu pemilihan nilai k (jumlah data/tetangga terdekat) [10].

2.2 Regresi logistik biner

Regresi logistik biner adalah analisis regresi yang digunakan pada saat peubah respon berskala kategori biner [11]. Model regresi logistik biner dibentuk dengan menyatakan nilai $E\ (Y = 1|x)$ sebagai $\pi(x)$, dimana $\pi(x)$ dinotasikan sebagai berikut [11]:

$$\pi(x) = \left[\frac{\exp\ (g(x))}{1 + \exp\ (g(x))}\right]$$

dengan $g(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, dimana

β_0 = konstanta
β_i = koefisien regresi logistic biner
i = $1, 2, \ldots, p$
p = banyak peubah penjelas.

Fungsi regresi di atas berbentuk non linier sehingga untuk membuatnya menjadi fungsi linier dilakukan transformasi logit sebagai berikut [12]:

$$\text{logit}[\pi(x)] = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = g(x).$$

2.3 Pohon Klasifikasi

Pohon klasifikasi adalah salah satu metode analysis statistika yang menerapkan penyekatan rekursif biner [13]. Pohon klasifikasi merumuskan aturan-aturan dalam penyekatan setiap simpul, penetapan simpul akhir dan penentuan nilai dugaan respon bagi setiap simpul akhir [14]. Dimana pemilahan yang dilakukan adalah memilah (mengkalsifikasikan) peubah tak bebas y skala kategorik biner berdasarkan peubah-peubah bebas x berjenis kategorik, kontinu ataupun kombinasi keduanya [14]. Untuk menentukan peubah bebas terbaik yang membedakan peubah y yaitu menghitung information gain dari setiap peubah bebas. Nilai *information gain* dapat diperoleh dengan mengetahui parameter lain yang disebut *entropy* yang mempengaruhi nilai gain. Entropy merupakan ukuran ketidakhomogenan kumpulan data [4].

$$\text{Entropy}(S) = \sum_i^c -p_i \log 2\, p_i.$$

dimana

c = jumlah nilai yang ada pada atribut target (jumlah kelas)
p_i = jumlah sampel pada kelas i.

Data yang heterogen dapat ditttunjukkan dengan entropi yang lebih besar dan sebaliknya data yang homogen akan memiliki entropi yang lebih kecil. Ketika kita sudah mendapatkan nilai *entropy*, maka langkah selanjutnya adalah melakukan perhitungan terhadap *information gain.* Nilai *gain* yang paling tinggi akan menjadi akar pertama Berdasarkan perhitungan matematis *information gain* dari suatu atribut A dapat diformulasikan sebagai berikut [4]:

$$\text{Gain}\ (S, A) = \text{Entropy}(S) - \sum_{\text{vevsiustion}(A)} \frac{|S_v|}{S} \text{entropy}(S_v)$$

dimana,

A	= atribut
V	= menyatakan suatu nilai yang mungkin untuk atribut A
Values (A)	= himpunan nilai-nilai yang mungkin untuk atribut A
|Sv|	= jumlah smpel untuk nilai v
|S|	= jumlah seluruh sampel data
Entropy (S)	= entropy untuk asmpel-sampel yang memiliki nilai v

Proses penyekatan terhadap simpul dilakukan secara berulang sampai ditemukan salah satu dari dua hal berikut:

1) Respon di semua simpul sudah homogen nilainya.
2) Jumlah objek di dalam simpul sudah terlalu sedikit untuk menghasilkan pemisahan yang memuaskan.

Dari proses penyekatan terbentuk pohon dengan berbagai ukuran. Pohon berukuran besar, biasanya menandakan pendugaan respon cenderung lebih tepat, namun sulit diinterpretasi. Sebaliknya saat pohon berukuran kecil, pohon mudah diinterpretasi namun pendugaan respon cenderung tidak tepat. Keseimbangan antara ukuran pohon dan ketepatan pendugaan respon adalah ciri khas yang dimiliki oleh pohon terbaik.

Pada software R, *minsplit* merupakan salah satu *hyperparameter* pada algoritma pohon klasifikasi yang ditentukan sebelum membuat model. *Minsplit* merupakan opsi untuk menentukan berapa ukuran node minimal untuk melakukan pemisahan pada proses penyekatan.

Langkah terakhir, dilakukan prediksi pada *testing data* untuk melihat akurasi klasifikasi pada masing-masing metode klasifikasi. Setelah itu menentukan metode klasifikasi terbaik berdasarkan tingkat akurasi tertinggi. Dalam rangka mengukur akurasi algoritma klasifikasi, penulis menggunakan *confusion matrix*. Metode ini menggunakan tabel matriks yang ditunjukkan pada Tabel 2 dimana jika data set hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negative [15].

Tabel 2. *Confusion Matrix*

Klasifikasi yang Benar	Diklasifikasikan sebagai	
	+	-
+	True positives	False negatives
-	False positives	True negatives

Setelah data uji dimasukkan ke dalam *confusion matrix*, selanjutnya dilakukan perhitungan untuk mengetahui nilai *sensitivity*, *specificity*, *precision* dan *accuracy*. *Sensitivity* digunakan untuk membandingkan jumlah data yang benar dalam pengklasifikasian (*true positives*) terhadap jumlah data yang masuk kelas positif sedangkan *specificity* adalah perbandingan jumlah data yang salah dalam pengklasifikasian (*true negatives*) terhadap jumlah data yang masuk kelas negatif. Ke empat nilai ini dihitung menggunakan persamaan berikut [4].

$$\text{sensitivity} = \frac{TP}{P},$$

$$\text{specificity} = \frac{TN}{N},$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{accuracy} = \text{sensitivity}\frac{P}{P+N} + \text{specificity}\frac{N}{P+N}$$

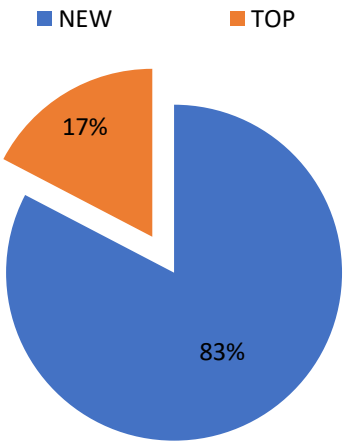
Keterangan:

TP = jumlah *true positives*
TN = jumlah *true negatives*
P = jumlah data yang masuk kelas positif
N = jumlah data yang masuk kelas negatif
FP = jumlah *false positives.*

3. Hasil dan Pembahasan

3.1 Deskripsi Data

Kredit Tanpa Agunan (KTA) Bank X merupakan layanan kredit dimana debitur tidak perlu memberikan jaminan atau agunan. *Top-Up* adalah fitur untuk meningkatkan jumlah pinjaman sebelum pinjaman periode sebelumnya sudah terbayar lunas. Dibawah ini akan disajikan gambaran umum tentang debitur KTA Bank X *Top Up* dan tidak dapat dilihat pada Gambar 1.



Gambar 1. Gambaran umum tentang debitur KTA Bank X

Ada sebanyak 83% data NEW, sedangkan data TOP hanya 17%. Data seperti ini disebut data yang tidak seimbang. Data tidak seimbang merupakan kondisi dimana terdapat perbedaan jumlah pengamatan yang sangat jauh berbeda antara satu kelas dibandingkan kelas lainnya. Selanjutnya akan dilakukan proses pembuatan model pada data dengan 13 peubah penjelas dengan jumlah data *training* sebanyak 80000 observasi, menggunakan metode *k-NN*, Regresi Logistik biner dan Pohon Klasifikasi. Model dibuat dengan bantuan *software R.*

3.2 Analisis Pemodelan k-NN

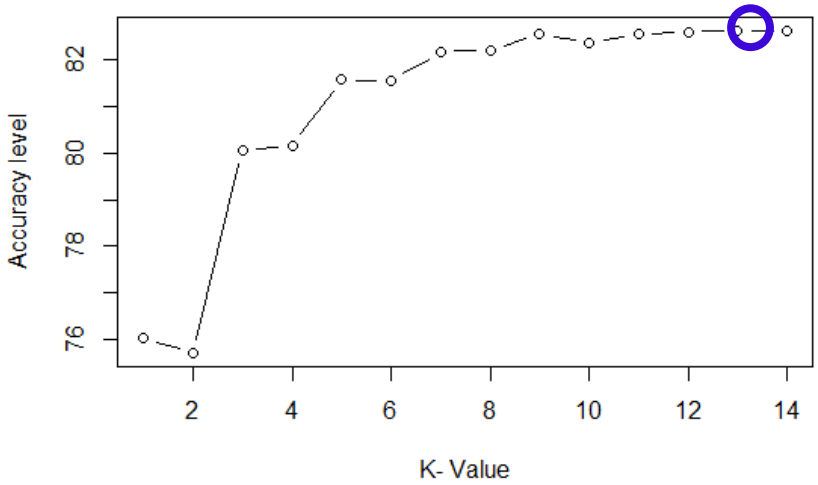
Dalam penggunaan *k-NN*, satuan data yang berbeda sangat mempengaruhi hasil analisis. Oleh karena itu, sebelum pengklasifikasian menggunnakann *k-NN* ada baiknya melakukan standarisasi atau transformasi data. Pada penelitian ini penulis menggunakan *min-max normalization* karena data terdiri dari skala pengukuran numerik dan kategorik. Setelah data di transformasi menggunakan *min-max normalization*, nilai tiap peubah yang di gunakan pada penelitian berkisar antara 0-1.

Selanjutnya, sebelum membuat model *k-NN* terlebih dahulu menentukan nilai k. Dalam menentukan k optimum penulis menggunakan confusion matrix untuk menghitung akurasi setiap k

yang digunakan. Dimulai dengan k=1 sampai k=50 selanjutnya diukur akurasinya untuk setiap k. Dari ke-50 k pilih akurasi optimum. K dengan akurasi optimum selanjutnya digunakan sebagai k optimum. Nilai-nilai akurasi untuk setiap k di jelaskan pada Tabel 3 dan Gambar 2.

Tabel 3. *Confusion Matrix*

K	Akurasi	K	Akurasi	K	Akurasi	K	Akurasi	K	Akurasi
1	76.05000	11	82.53571	21	82.17143	31	80.05000	41	82.62143
2	75.71429	12	82.59286	22	82.20714	32	80.15714	42	82.61429
3	80.05000	13	82.62143	23	82.53571	33	81.56429	43	76.05000
4	80.15714	14	82.61429	24	82.35714	34	81.53571	44	75.71429
5	81.56429	15	76.05000	25	82.53571	35	82.17143	45	80.05000
6	81.53571	16	75.71429	26	82.59286	36	82.20714	46	80.15714
7	82.17143	17	80.05000	27	82.62143	37	82.53571	47	81.56429
8	82.20714	18	80.15714	28	82.61429	38	82.35714	48	81.53571
9	82.53571	19	81.56429	29	76.05000	39	82.53571	49	82.17143
10	82.35714	20	81.53571	30	75.71429	40	82.59286	50	82.20714



Gambar 2. Grafik nilai akurasi untuk $k = 1$ sampai $k = 14$

Pada Tabel 3 dan Gambar 3 terlihat bahwa k yang optimum adalah k=13 dengan akurasi 82.62%. Selanjutnya akan dilakukan pemodelan *k-NN* dengan $k = 13$ pada data *training* dan menghitung kinerja model tersebut menggunakan *confusion matrix* sehingga diperoleh hasil yang ditunjukkan pada Tabel 4.

Tabel 4. Hasil klasifikasi model *k-NN* dengan $k = 13$ pada data training

Keadaan Sesungguhnya	Prediksi	
	TOP	NEW
TOP	2041	11809
NEW	1118	65032
Akurasi	83.84%	

Berdasarkan Tabel 4, model *k-NN* dengan $k = 13$ berhasil mengklasifikasi dengan benar debitur yang berpotensi *top-up* pada data *training* sebanyak 2.041 debitur atau sebesar 15% (nilai *sensitivity)* dan mengklasifikasi dengan benar debitur yang tidak berpotensi *top-up* pada data *training* sebanyak 65.032 debitur atau sebesar 98% (nilai *specificity)*. Jika dicermati antara nilai *sensitivity* dan *specificity* terdapat perbedaan yang sangat timpang dimana nilai *sensitivity* sangat kecil. Artinya

model *k-NN* masih kurang baik dalam memprediksi dengan benar debitur yang berpotensi *top-up.* Sebanyak 85% debitur yang seharusnya berpotensi *top-up* diprediksi tidak berpotensi *top*-up. Hal ini bisa saja terjadi karena adanya ketidak seimbangan data pada variabel status *upselling.* Sehingga walaupun secara keseluruhan akurasi klasifikasi debitur menggunakan data *training* sebesar 83% sudah cukup baik, namun sebenanrnya model *k-NN* masih belum cukup handal dalam mengatasi ketidakseimbangan data.

Selanjutnya model *k-NN* yang dibentuk pada data *training* digunakan untuk memprediksi status *upselling* debitur pada data *testing.* Hasil prediksi ditunjukkan pada Tabel 5.

Tabel 5. Hasil klasifikasi model *k-NN* pada data testing

Keadaan Sesungguhnya	Prediksi	
	TOP	NEW
TOP	11301	244
NEW	2188	267
Akurasi	82.62%	

3.3 Analisis Pemodelan Regresi Logistik Biner

Pemodelan menggunakan metode regresi logistik biner dimulai dengan melakukan pemodelan menggunakan data *training* untuk mendapatkan model regresi logistik terbaik. Berikut disajikan pada Tabel 6 hasil pendugaan parameter dan *p-value* dari masing-masing peubah.

Tabel 6. Pendugaan parameter menggunakan regresi logistik biner pada data *training*

Koefisien	Estimate	Std. Error	Z	Value	$Pr(>\|Z\|)$
(Intercept)	-2.38E+00	9.78E-02	-24.362	2.00E-16	***
JK 1	-1.31E-02	2.71E-02	-0.482	0.629619	
Usia	1.00E-03	1.99E-03	0.502	0.615979	
Status Nikah 1	-7.00E-02	3.35E-02	-2.091	0.036489	*
Status Nikah 2	4.81E-03	8.63E-02	0.056	0.955544	
Tanggungan	-3.20E-02	1.51E-02	-2.12	0.033985	*
Pendidikan S1	-2.29E-01	3.16E-02	-7.259	3.89E-13	***
Pendidikan S2/S3	-6.05E-01	8.83E-02	-6.846	7.57E-12	***
Pendidikan SD	2.83E-01	1.75E-01	1.619	0.105531	
Pendidikan SMA	1.45E-01	2.79E-02	5.187	2.14E-07	***
Pendidikan SMP	3.43E-01	9.80E-02	3.498	0.000469	***
Wilayah Bandung	-1.88E-01	4.80E-02	-3.919	8.90E-05	***
Wilayah Denpasar	-2.28E-02	7.71E-02	-0.295	0.767956	
Wilayah Jakarta 3	-1.19E-01	4.64E-02	-2.562	0.010415	*
Wilayah Jakarta 4	-1.46E-01	4.65E-02	-3.136	0.001713	**
Wilayah Jakarta 5	5.88E-04	4.44E-02	0.013	0.989436	
Wilayah Jayapura	-3.07E-01	7.41E-02	-4.144	3.42E-05	***
Wilayah Makassar	-1.54E-01	7.04E-02	-2.188	0.028676	*
Wilayah Medan	-9.68E-02	4.53E-02	-2.135	0.032797	*
Wilayah Palembang	-3.67E-01	5.57E-02	-6.585	4.53E-11	***
Wilayah Semarang	-1.67E-02	5.75E-02	-0.291	0.770792	
Wilayah Surabaya	1.69E-01	4.62E-02	3.661	0.000252	***
Pekerjaan Pegawai Swasta	7.80E-01	3.85E-02	20.232	2.00E-16	***
Pekerjaan PNS	1.71E-01	5.43E-02	3.144	0.001667	**

Koefisien	Estimate	Std. Error	Z	Value	$Pr(>\|Z\|)$
Lama Bekerja	-8.30E-03	2.16E-03	-3.851	0.000118	***
Penghasilan	-1.42E-07	6.09E-09	-23.242	2.00E-16	***
DBR	-1.04E+00	9.90E-02	-10.54	2.00E-16	***
Waktu Kredit	6.90E-03	5.08E-04	13.589	2.00E-16	***
Limit Kredit	1.47E-08	3.06E-10	48.046	2.00E-16	***
Gaji Tanggungan	-2.05E-08	8.25E-09	-2.485	0.01297	*

Berdasarkan Tabel 6, terdapat beberapa peubah yang tidak signifikan (dikatakan signifikan jika nilai $p_{\text{value}} < 0.05$) yaitu peubah jenis kelamin dan usia, yang artinya kedua peubah ini belum cukup bukti untuk menyatakan ada pengaruh terhadap status *upselling*. Sehingga dilakukan pemodelan sekali lagi dengan membuang peubah yang tidak signifikan. Adapun hasil pendugaan parameter dengan membuang peubah usia dan jenis kelamin di sajikan pada Tabel 7.

Tabel 7. Pendugaan parameter menggunakan regresi logistik biner pada data *training* tanpa peubah Jenis Kelamin dan Usia

Koefisien	Estimate	Std. Error	Z	Value	$Pr(>\|Z\|)$
(Intercept)	-2.38E+00	9.78E-02	-24.362	2.00E-16	***
Status Nikah 1	-6.59E-02	3.28E-02	-2.008	0.044694	*
Status Nikah 2	1.44E-02	8.50E-02	0.17	0.865292	
Tanggungan	-3.09E-02	1.45E-02	-2.132	0.03302	*
Pendidikan S1	-2.28E-01	3.15E-02	-7.241	4.46E-13	***
Pendidikan S2/S3	-6.02E-01	8.82E-02	-6.824	8.83E-12	***
Pendidikan SD	2.85E-01	1.75E-01	1.633	0.102447	
Pendidikan SMA	1.43E-01	2.77E-02	5.164	2.41E-07	***
Pendidikan SMP	3.45E-01	9.79E-02	3.518	0.000435	***
Wilayah Bandung	-1.88E-01	4.79E-02	-3.921	8.80E-05	***
Wilayah Denpasar	-2.12E-02	7.70E-02	-0.275	0.783018	
Wilayah Jakarta 3	-1.18E-01	4.63E-02	-2.557	0.01055	*
Wilayah Jakarta 4	-1.45E-01	4.62E-02	-3.142	0.001679	**
Wilayah Jakarta 5	1.79E-03	4.42E-02	0.04	0.967765	
Wilayah Jayapura	-3.05E-01	7.40E-02	-4.126	3.69E-05	***
Wilayah Makassar	-1.53E-01	7.04E-02	-2.175	0.029643	*
Wilayah Medan	-9.51E-02	4.53E-02	-2.103	0.035491	*
Wilayah Palembang	-3.66E-01	5.57E-02	-6.575	4.88E-11	***
Wilayah Semarang	-1.39E-02	5.72E-02	-0.243	0.80788	
Wilayah Surabaya	1.70E-01	4.62E-02	3.679	0.000234	***
Pekerjaan Pegawai Swasta	7.79E-01	3.85E-02	20.226	2.00E-16	***
Pekerjaan PNS	1.73E-01	5.41E-02	3.195	0.0014	**
Lama Bekerja	-7.66E-03	1.79E-03	-4.273	1.93E-05	***
Penghasilan	-1.42E-07	6.08E-09	-23.285	2.00E-16	***
DBR	-1.04E+00	9.90E-02	-10.532	2.00E-16	***
Waktu Kredit	6.86E-03	5.03E-04	13.636	2.00E-16	***
Limit Kredit	1.47E-08	3.06E-10	48.046	2.00E-16	***
Gaji Tanggungan	-2.02E-08	8.24E-09	-2.456	0.014049	*

Berdasarkan Tabel 7 terlihat bahwa semua peubah sudah signifikan, sehingga dilanjutkan dengan menghitung kinerja model regresi logistik biner tanpa peubah usia dan jenis kelamin pada

data *training* menggunakan *confusion matrix* sehingga diperoleh hasil yang ditunjukkan pada Tabel 8.

Tabel 8. Hasil klasifikasi model regresi logistik biner tanpa peubah usia dan jenis kelamin pada data *training*

Keadaan Sesungguhnya	Prediksi	
	TOP	NEW
TOP	1988	11862
NEW	1379	64771
Akurasi	83.45%	

Berdasarkan Tabel 8, model regresi logistik biner tanpa peubah usia dan jenis kelamin berhasil mengklasifikasi dengan benar debitur yang berpotensi *top-up* pada data *training* sebanyak 1.988 debitur atau sebesar 14% (nilai *sensitivity)* dan mengklasifikasi dengan benar debitur yang tidak berpotensi *top-up* pada data *training* sebanyak 64.771 debitur atau sebesar 98% (nilai *specificity)*. Jika dicermati dengan menggunakan model ini antara nilai *sensitivity* dan *specificity* terdapat perbedaan yang sangat timpang dimana nilai *sensitivity* sangat kecil. Hal ini serupa dengan pemodelan *k-NN*, walaupun secara keseluruhan akurasi klasifikasi debitur menggunakan data *training* sebesar 83.45% sudah cukup baik, namun sebenanrnya model regresi logistik biner yang dibentuk masih belum cukup handal dalam mengatasi ketidakseimbangan data. Selanjutnya model regresi logistik biner tanpa peubah usia dan jenis kelamin digunakan untuk memprediksi status *upselling* debitur pada data *testing.* Hasil prediksi ditunjukkan pada Tabel 9.

Tabel 9. Hasil klasifikasi model regresi logistik biner tanpa peubah usia dan jenis kelamin pada data *testing*

Keadaan Sesungguhnya	Prediksi	
	TOP	NEW
TOP	331	2124
NEW	248	11297
Akurasi	83.06%	

3.4 Analisis Pemodelan Pohon Klasifikasi

Pemodelan menggunakan metode pohon klasifikasi dimulai dengan melakukan pemodelan menggunakan data *training* dengan *minsplit* 300. Berikut disajikan kinerja model pohon klasifikasi dengan minsplit 300 pada data *training* menggunakan *confusion matrix* sehingga diperoleh hasil yang ditunjukkan pada Tabel 10.

Tabel 10. Hasil klasifikasi model pohon klasifikasi dengan minsplit 300 pada data *training*

Keadaan Sesungguhnya	Prediksi	
	TOP	NEW
TOP	7463	2858
NEW	6387	63292
Akurasi	88.44%	

Berdasarkan Tabel 10, model pohon klasifikasi berhasil mengklasifikasi dengan benar debitur yang berpotensi *top-up* pada data *training* sebanyak 7.463 debitur atau sebesar 72% (nilai *sensitivity)* dan mengklasifikasi dengan benar debitur yang tidak berpotensi *top-up* pada data *training* sebanyak 63.292 debitur atau sebesar 91% (nilai *specificity)*. Jika dicermati dengan menggunakan model ini antara nilai *sensitivity* dan *specificity* tidak terdapat perbedaan yang sangat timpang, sehingga bisa

dikatakan model pohon klasifikasi cukup handal dalam mengatasi ketidakseimbangan data dengan akurasi yang cukup tinggi yaitu 88.44% yang lebih besar disbanding kedua model sebelumnya.

Selanjutnya model pohon klasifikasi dengan *minsplit* 300 digunakan untuk memprediksi status *upselling* debitur pada data *testing.* Hasil prediksi ditunjukkan pada Tabel 11. Berdasarkan metode Pohon Klasifikasi, peubah peubah yang digunakan dalam aturan di Pohon Klasifikasi adalah DBR, lama bekerja seorang debitur, limit kredit, jenis pekerjaan debitur, jumlah penghasilan debitur, wilayah tempat tinggal debitur serta jangka waktu kredit debitur selama 1 bulan.

Tabel 11. Hasil klasifikasi model pohon klasifikasi dengan minsplit 300 pada data *testing*

Keadaan Sesungguhnya	Prediksi	
	TOP	NEW
TOP	1255	525
NEW	1200	11020
Akurasi	87.68%	

3.5 Pemilihan metode terbaik

Setelah diterapkan metode *k-NN*, Regresi Logistik biner dan Pohon Klasifikasi pada data KTA Bank X, berikut disajikan rangkuman hasil akurasi dari ketiga metode yang ditunjukkan pada Tabel 12.

Tabel 12. Rangkuman Kinerja klasifikasi dari 3 metode

Metode Klasifikasi	Akurasi	
	Data Training	Data Testing
KNN	83.84	82.62
Regresi Logistik Biner	83.45	83.06
Pohon Klasifikasi	88.44	87.68

Berdasarkan Tabel 12, akurasi tertinggi pada data *training* dan data *testing* adalah model dengan menggunakan metode Pohon Klasifikasi. Oleh karena itu, metode terbaik pada penelitian ini dalam memprediksi debitur KTA (kredit tanpa agunan) di Bank X yang berpotensi melakukan *top-up* kredit atau tidak adalah metode pohon klasifikasi.

4. Kesimpulan

Berdasarkan hasil yang diperoleh, dengan melihat perbandingan nilai *accuracy*, metode Pohon Klasifikasi menghasilkan nilai *accuracy* yang paling tinggi dibandingkan kedua metode lainnya. Dengan demikian, untuk penelitian ini metode Pohon Klasifikasi menghasilkan prediksi lebih baik yaitu 87,68%. Berdasarkan metode Pohon Klasifikasi, peubah peubah yang digunakan dalam aturan di Pohon Klasifikasi adalah DBR (*Debt Burden Ratio)*, lama bekerja seorang debitur, limit kredit, jenis pekerjaan debitur, jumlah penghasilan debitur, wilayah tempat tinggal debitur serta jangka waktu kredit debitur selama 1 bulan.

Ucapan Terima Kasih

Penelitian ini didanai melalui dana DIPA PNBP Universitas Tanjungpura (SP DIPA-023.17.2.677517/2022) tahun anggaran 2022.

Referensi

[1] Kasmir, *Bank dan Lembaga Keuangan Lainnya*. Jakarta: PT. Raja Grafindo Persada, 2002.

[2] R. Odegua, "Predicting Bank Loan Default with Extreme Gradient Boosting," *arXiv - CS – Statistical Finance; Machine Learning*. 2020.

[3] D. A. Salazar, J. L. Velez, J. C. Salazar, "Comparison between svm and logistic regression : which one is better to discriminate?" *Revista Colombiana de Estadistica*, 35(SPE2), pp. 223-237, June. 2012.

[4] J. Han and M. Kamber, *Data Mining Concept and Tehniques*. San Fransisco: Morgan Kauffman, 2006.

[5] Y. Maldini, A. M. Siregar dan T. A. Mudzakir, "Perbandingan Algoritma C4.5 dan KNN Untuk Menentukan Pemberian Kredit Bagi Nasabah Koperasi," *Scientific Student Journal for Information, Technology and Science.*, vol. 2, no.1, pp. 31-38, Januari. 2021.

[6] S. Sreesouthry, A. Ayubkhan, M. M. Rizwan, D. Lokesh dan K. P. Raj, "Loan Prediction Using Logistic Regression in Machine Learning," *Annals of The Romanian Society for Cell Biology.*, vol. 25, no. 4, pp. 2790 – 2794, 2021.

[7] M. Madaan, A. Kumar, C. Keshri , R. Jain dan P. Nagrath, "Loan default prediction using decision trees and random forest: A comparative study," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, pp. 012042, 2020.

[8] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining*. New York: CRC Press, 2009.

[9] D. T. Larose, *Discovering Knowledge in Data*. New Jersey: John Willey & Sons, Inc., 2005.

[10] B. Santoso, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis 1 edition.* Yogyakarta: Graha Ilmu, 2007.

[11] D.W. Hosmer and Lemeshow S, *Applied Logistic Regression, Second Edition.* New York: John Wiley and Sons, 2002.

[12] A. Agresti, *Categorical Data Analysis.* New Jersey: John Wiley and Sons, 1990.

[13] R. J. Lewis, "An Introduction to Classification and Regression Trees (CART) Analysis," in The Annual Meeting of the Society for Academic Emergency Medicine, California, UCLA Medical Center, 2000.

[14] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*. New York: Chapman Hall, 1993

[15] Bramer and Max, *Principles of Data Mining*. London: Springer, 2007