UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2 Nomor 4 Tahun 2024 Halaman 422-429.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Implementation of CART Method with SMOTE for Household Poverty Classification in Mentawai Islands 2023

Rheizma Dewi Adiningtiyas, Admi Salma*, Syafriandi, Fadhilah Fitri
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: admisalma1@fmipa.unp.ac.id

Submitted : 08 Agustus 2024
Revised : 15 Agustus 2024
Accepted : 11 November 2024

ABSTRACT

Poverty is a condition in which individuals or groups are unable to fulfill their basic needs due to economic pressure or limited resources. The Classification and Regression Trees (CART) method is a classification technique in the form of a classification tree, which describes the relationship between independent and dependent variables. Data imbalance can lead to low sensitivity values and area under curve (AUC) values. One method that can overcome unbalanced data is to perform Synthetic Minority Oversampling Technique (SMOTE). SMOTE is a technique with the addition of artificial data in the minority class at a stage before analyzing the data. The purpose of this research is to compare the model without and with SMOTE in CART method. The use of SMOTE is applied to balance the amount of data on each poor household. The accuracy value of the method without SMOTE is 89% while with the SMOTE method is 79%. However, the sensitivity value has increased by 80%. Meanwhile, the AUC value in the CART method with SMOTE increased by 31%. So in this study it can be concluded that CART classification analysis with SMOTE is able to provide better performance compared to CART classification analysis without SMOTE.

Keywords: *CART, Mentawai, SMOTE*

I. PENDAHULUAN

Pemerintah pusat dan daerah terus memberikan perhatian khusus pada masalah kemiskinan karena masalah ini masih menjadi masalah yang rumit bagi negara-negara berkembang seperti Indonesia. Kemiskinan dapat didefinsikan ketika seseorang atau masyarakat tidak dapat memenuhi kebutuhan dasar mereka karena kesulitan keuangan atau sumber daya yang langka. Salah satu hal yang menyebabkan kemiskinan adalah perbedaan distribusi pendapatan antara kelompok berpenghasilan tinggi dan rendah. Meningkatnya kemiskinan juga dapat diakibatkan oleh tingginya jumlah individu dan kelompok yang hidup di bawah garis kemiskinan. Berbagai masalah kemiskinan di Indonesia bersumber dari ketidakmerataan distribusi pendapatan antara kelompok berpenghasilan tinggi dan rendah.

Jika membandingkan daerah lain di Sumatera Barat, Kepulauan Mentawai memiliki tingkat kemiskinan yang paling tinggi, sehingga masalah kemiskinan di sana menjadi perhatian serius. Pengentasan kemiskinan di Kabupaten Kepulauan Mentawai tidak hanya meningkatkan taraf hidup masyarakat yang terkena dampak langsung, namun juga mendukung pertumbuhan sosial, ekonomi, dan lingkungan dalam jangka panjang.

Rumah tangga yang merupakan unit terkecil ini adalah tempat pertama kali kemiskinan terlihat. Menurut BPS (2021), sebuah rumah tangga dikatakan miskin jika rata-rata pengeluaran konsumsi per kapita per bulan berada di bawah garis kemiskinan, dan sebaliknya. Jika rata-rata pengeluaran konsumsi bulanan dan pendapatan per kapita suatu rumah tangga melebihi kriteria kemiskinan, maka rumah tangga tersebut tidak dapat diklasifikasikan sebagai rumah tangga miskin. Analisis klasifikasi dapat digunakan untuk mengidentifikasi ciri-ciri keluarga yang miskin dan yang tidak miskin. Dalam penelitian ini, analisis *Classification and Regression Tree (CART)* digunakan sebagai teknik klasifikasi.

CART merupakan salah satu dari 10 teknik data mining teratas. Faktor-faktor yang paling penting dan interaksinya dapat dipilih dengan menggunakan pendekatan CART untuk mengidentifikasi variabel respon. Kemampuan teknik CART dalam menganalisis algoritma sesuai dengan waktu tempuh dan tingkat akurasi yang diperlukan telah diakui. Sartono dan Syafitri (2010) menjelaskan bahwa CART dapat mempermudah dalam memahami hasil analisis dan menghasilkan kesimpulan dengan tingkat kesalahan yang rendah. Akurasi CART yang lebih besar, yaitu 95,2%, telah

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/232

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

diamati pada penelitian sebelumnya oleh Ispriyanti dkk. (2019) yang membandingkan *Classification and Regression Tree* (CART) dengan *Chisquare Automatic Interaction Detection* (CHAID).

Klasifikasi akan menjadi kurang akurat karena ada ketidakseimbangan dalam jumlah data untuk variabel dependen. Ketika variabel dependen tidak seimbang, ini menunjukkan bahwa kelas mayoritas adalah kelas yang memiliki lebih banyak data. Di sisi lain, kelas minoritas adalah kelas yang memiliki data yang lebih sedikit. Ketika ada ketidakseimbangan dalam data, klasifikasi sering kali mengklasifikasikan data dari kelas mayoritas sementara mengabaikan data dari kelas minoritas, yang menurunkan nilai akurasi pada kelas minoritas (Chawla dkk., 2002). Oleh karena itu, teknik *Synthetic Minority Oversampling Technique* (SMOTE) berkontribusi dalam memunculkan data palsu pada kelas minoritas sebelum dianalisis menggunakan CART, yang bertujuan untuk meningkatkan nilai akurasi dari perkiraan pohon klasifikasi.

Untuk memperbaiki data yang tidak merata dan meningkatkan rasio data mayor dan minor, algoritma SMOTE menciptakan data baru di kelas minoritas. Penelitian ini berbeda dengan penelitian lainnya karena menggunakan data statistik kemiskinan rumah tangga di Kabupaten Kepulauan Mentawai. Klaim ini memandu tujuan penelitian, yaitu menerapkan algoritma SMOTE dan model klasifikasi CART untuk menyeimbangkan data kelas minoritas dalam rangka mengklasifikasikan kemiskinan rumah tangga di Kabupaten Kepulauan Mentawai pada tahun 2023.

II. METODE PENELITIAN

Data yang digunakan dalam penelitian ini berasal data Survei Sosial Ekonomi Nasional (SUSENAS) Kabupaten Kepulauan Mentawai Tahun 2023 dengan jumlah responden sebanyak 535 kepala keluarga. Data dari BPS Provinsi Sumatera Barat. Perangkat lunak yang digunakan adalah *Rstudio*. Variabel dependen dan variabel independen adalah dua kategori variabel yang digunakan dalam penelitian ini. Variabel dependen yang digunakan terdiri dari keluarga yang diklasifikasikan sebagai keluarga miskin (nol) dan tidak miskin (satu). Mengenai lima belas variabel independen dalam penelitian ini terdiri dari Jumlah Anggota Keluarga (X_1), Usia Kepala Keluarga (X_2), Pendidikan Kepala keluarga (X_3), Pekerjaan Kepala Keluarga (X_4), Status Kepemilikan Rumah (X_5), Luas Lantai (m^2) (X_6), Jenis Lantai (X_7), Jenis Dinding (X_8), Jenis Atap (X_9), Jenis Penerangan (X_{10}), Sumber Air Minum (X_{11}), Bahan Bakar Untuk Memasak (X_{12}), Fasilitas Buang Air Besar (X_{13}), Memiliki Jaminan Kesehatan (X_{14}), Menerima Bantuan PKH (Program Keluarga Harapan) dan sejenisnya (X_{15}). Metode yang digunakan adalah CART dengan langkah-langkah sebagai berikut.

1. Pembuatan Pohon Klasifikasi
   a. Pemilihan pemilah

   Pembuatan pohon klasifikasi dilakukan dengan mencari pengklasifikasi dari setiap *node* yang dapat mengurangi tingkat impuritas yang paling tinggi. Impuritas suatu simpul dapat diukur melalui nilai impuritasnya. Menurut Breiman dkk (1984), semakin besar nilai impuritas suatu simpul, maka semakin heterogen simpul tersebut. Probabilitas pengamatan yang masuk pada simpul kanan dan kiri dihitung dengan rumus sebagai berikut (Mardiani, 2012):

$$P_L = \frac{calon\ simpul\ kiri}{data\ training} \quad (1)$$

$$P_R = \frac{calon\ simpul\ kanan}{data\ training} \quad (2)$$

   Nilai impuritas pada simpul t didefinisikan sebagai berikut:

$$i(t) = 1 - \sum_{j=1} p^2(j|t) \quad (3)$$

(Breiman dkk, 1984)

   $P\ (j|t)$ adalah adalah probabilitas kelas j pada simpul t. Penyekatan (s) pada simpul t adalah sebagai penurunan impuritas, sebagai berikut:

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \quad (4)$$

   Dimana :
   $i(t)$: Fungsi impuritas pada simpul
   $i(t_R)$: Fungsi impuritas pada simpul anak kanan
   $i(t_L)$: Fungsi impuritas pada simpul anak kiri
   P_R : Probabilitas pengamatan pada simpul kanan
   P_L : Probabilitas pengamatan pada simpul kiri

(Breiman dkk, 1984)

   Nilai *goodness of split* terbesar yang merupakan pemilah utama yang akan menjadi *root node*.

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/232

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

b. Penentuan Simpul Terminal

Pemilahan menentukan apakah sebuah simpul merupakan simpul terminal atau bukan berdasarkan dua faktor, yaitu jika tingkat kedalaman pohon maksimal dan jumlah pengamatan kurang dari atau sama dengan lima ($n < 5$). Pengembangan pohon akan berakhir setelah hal ini selesai.

c. Penandaan Label Kelas

Dengan menggunakan aturan bilangan terbesar, label kelas pada simpul terminal ditandai sebagai berikut:

$$P\ (j_0|t) = {}^{max}_{j}\ P(j|t) = {}^{max}_{j}\ \frac{N_j(t)}{N(t)} \quad (5)$$

Dimana :

$P\ (j|t)$: Probabilitas kelas j pada node t

$N_j(t)$: Jumlah pengamatan kelas j pada node t

$N(t)$: Jumlah pengamatan pada node t

(Pratiwi dan Zain, 2014)

2. Pemangkasan Pohon Klasifikasi

Pemangkasan pohon dilakukan berdasarkan ukuran kompleksitas biaya terkecil, dimana cabang yang dipangkas adalah cabang yang memiliki nilai paling rendah dengan menggunakan rumus sebagai berikut, sehingga diperoleh ukuran pohon yang layak:

$$g_m(t) = \frac{R(t)-R(T_k)}{|T_k|-1} \quad (6)$$

Dimana :

$g_m(t)$: *Complexity parameter*

$R(t)$: Kesalahan pengklasifikasian pada node t

T_k : Subtree ke-k, dengan k= 1,2,...,n

$R(T_k)$: Kesalahan pengklasifikasian pada pohon

3. Penentuan Pohon Klasifikasi Optimal

Penentuan ini dilakukan berdasarkan penduga *cross validation* sebagai berikut:

$$R\left(T_t^{(v)}\right) = \frac{1}{N_v} \sum X\left(d^{(v)}\right) \quad (7)$$

Dimana $X\left(d^{(v)}\right)$ hasil pengklasifikasian dan N_v jumlah pengamatan dalam L_v.

SMOTE adalah sebuah pendekatan yang digunakan pada menyeimbangkan data sampel dari kelas yang terlalu tidak seimbang *(mayoritas)* dengan penekanan pada kelas negatif *(minoritas),* dengan tujuan meningkatkan kinerja metode klasifikasi. Teknik statistik nonparametrik (*K-Nearest Neighbor*), yang merupakan teknik statistik nonparametrik, merupakan metodologi yang digunakan dalam algoritma SMOTE. Untuk jumlah *k-nearest neighbor,* dapat ditentukan berdasarkan pertimbangan kegunaannya. Untuk pembuatan data buatan yang berskala kategorik menggunakan rumus *Value Difference Metric* (VDM). (Chawla, et al, 2002).

$$\Delta\ (A,B) = \sum_{i=1}^{N} \delta\ (V_{1i}, V_{2i}) \quad (8)$$

Dimana :

$\Delta\ (A,B)$: Jarak antara amatan A dengan amatan B

N : Banyaknya variabel independen

$\delta(V_{1i}, V_{2i})$: Jarak antara amatan A dan B untuk setiap variabel yang dihitung

Untuk menentukan jarak antar amatan A dan B untuk setiap variabel maka digunakan persamaan:

$$\delta(V_{1i}, V_{2i}) = \sum_{i=1}^{n} \left|\frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2}\right| \quad (9)$$

Dimana :

n : Banyaknya kategori pada variabel ke-i

C_{1i} : Banyaknya kategori ke-1 yang termasuk pada variabel ke-i

C_{2i} : Banyaknya kategori ke-2 yang termasuk pada variabel ke-i

C_1 : Banyaknya kategori ke-1 terjadi

C_2 : Banyaknya kategori ke-2 terjadi

Tiga perhitungan statistik dapat digunakan untuk mengukur tingkat akurasi model klasifikasi: akurasi, sensitivitas, dan spesifisitas. Rumus untuk spesifisitas, sensitivitas, dan akurasi adalah sebagai berikut. Tabel 1 merupakan alat ukur berbentuk matriks yang menunjukkan tingkat akurasi klasifikasi kelas dengan menggunakan algoritma yang digunakan (Pratiwi dan Zain, 2014). Nilai akurasi klasifikasi diperoleh dengan menggunakan persamaan berikut:

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/232

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+F} \quad (10)$$

$$\text{Sensitivitas} = \frac{TP}{TP+FN} \quad (11)$$

$$\text{Spesifisitas} = \frac{TN}{FP+TN} \quad (12)$$

Tabel 1. *Confusion Matrix*

Aktual	Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Kurva *Receiver Operating Characteristics* (ROC) adalah metode yang digunakan untuk memvisualisasikan dan memilih model klasifikasi terbaik berdasarkan kinerja. Pada kurva ROC, tingkat *True Positive* (TP) diplot pada sumbu Y sedangkan tingkat *False Positive* (FP) diplot pada sumbu X. ROC memiliki area yang disebut *Area Under Curve* (AUC) yang dapat digunakan untuk membandingkan kinerja beberapa model klasifikasi untuk menemukan model terbaik. Nilai AUC berkisar antara 0 hingga 1, dimana jika mendekati 1, maka dapat dikatakan bahwa model tersebut mampu mengklasifikasikan dengan baik.. Adapun perhitungan *AUC* didefinisikan sebagai berikut:

$$AUC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (14)$$

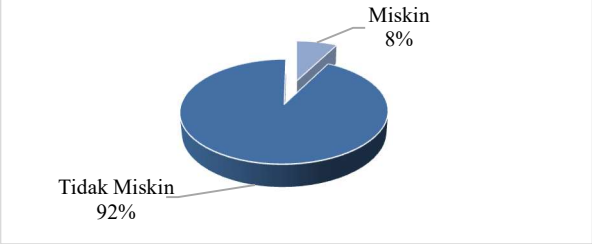
Nilai dalam pengklasifikasian AUC dibagi dalam beberapa kelompok yaitu sebagai berikut:

a. 0,90 – 1,00 = Klasifikasi sangat baik
b. 0,80 – 0,90 = Klasifikasi baik
c. 0,70 – 0,80 = Klasifikasi cukup
d. 0,60 – 0,70 = Klasifikasi buruk
e. 0,50 – 0.60 = Klasifikasi salah

(Sari dkk, 2020)

III. HASIL DAN PEMBAHASAN

Ketika ingin melakukan analisis data, yang harus dilakukan terlebih dahulu adalah melihat keseimbangan data yang akan digunakan, deskripsi data kemiskinan rumah tangga pada tahun 2023 disajikan seperti pada Gambar 1.



Gambar 1. Deskripsi Data Kemiskinan Rumah Tangga Tahun 2023

Gambar 1 memperlihatkan bahwa kelas dengan jumlah variabel dependen terbesar atau kelas utama merupakan 92% dari kelompok tidak miskin. Rumah tangga tidak miskin adalah individu-individu yang memiliki rata-rata pengeluaran konsumsi per kapita per bulan di atas garis kemiskinan. Rata-rata pengeluaran konsumsi per kapita per bulan masyarakat berada di bawah ambang batas kemiskinan yaitu sebesar 8%, menurut statistik minor.

Dengan demikian, untuk kelas minor atau kelompok miskin, penting untuk mencari data sintetis dengan menggunakan persamaan (8) berdasarkan tetangga terdekat dengan menggunakan jarak *Value Difference Metric* (VDM) pada persamaan (9). Untuk mendapatkan jumlah tanggungan yang sama dengan kelas utama, diperlukan 11 kali replikasi untuk kelas minor dari kelompok miskin, yang terdiri dari 41 keluarga yang memiliki tanggungan. Tetangga terdekat untuk masing-masing setelah 11 kali replikasi adalah 5 data dengan data baru. Tabel 1 menampilkan distribusi data setelah replikasi teknik SMOTE:

Tabel 2. Distribusi Data dengan Menggunakan SMOTE

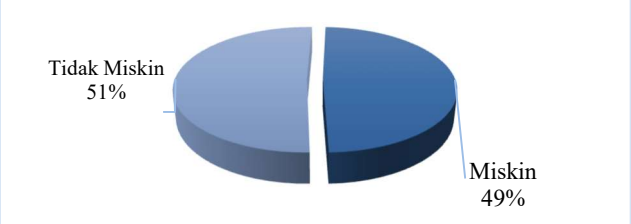
Sebelum SMOTE		**Sesudah SMOTE**		**Jumlah Replikasi**
Mayor	**Minor**	**Mayor**	**Minor**	
Tidak Miskin	Miskin	Tidak Miskin	Miskin	11 kali

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/232

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

494 (92%)	41 (8%)	494 (51%)	492 (49%)

Terlihat dari Tabel 2 bahwa akan terjadi peningkatan data dari 41 titik data awal menjadi 492 titik data yang mengindikasikan bahwa jumlah variabel masing-masing kelas seimbang. Jumlah data pada setiap variabel independen akan meningkat secara berurutan setelah jumlah data pada variabel dependen. Hal ini sebagai tambahan dari jumlah data pada variabel dependen itu sendiri. Hal ini dimaksudkan agar dengan menyeimbangkan jumlah anggota pada setiap jenis data pada variabel dependen, skenario *underfitting* dan *overfitting* dapat dihindari dan tingkat akurasi yang masuk akal dapat dihasilkan. Oleh karena itu, Gambar 2 menunjukkan jumlah anggota untuk kedua kelas pada variabel dependen setelah replikasi.



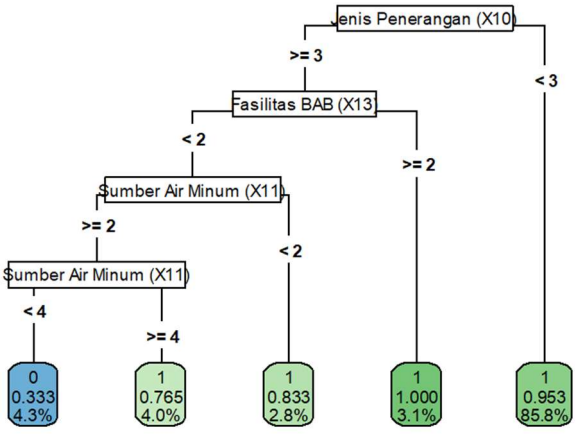
Gambar 2. Persentase Banyaknya Variabel Dependen Setelah SMOTE

Gambar 2 mengilustrasikan bagaimana setiap bentuk data variabel dependen memiliki persentase data yang seimbang. Jumlah data penelitian yang digunakan telah meningkat dari 41 data di awal menjadi 492 data. Untuk menyeimbangkan jumlah data pada kelas utama, replikasi SMOTE digunakan untuk membuat data sintetis. Dengan demikian, setelah direplikasi sebanyak 11 kali, jumlah rumah tangga miskin meningkat menjadi 492, atau 49% dari total keseluruhan, sementara jumlah data keluarga yang tidak termasuk dalam kelas utama miskin tetap berada di angka 494, atau 51%. Setelah mengumpulkan data yang cukup dengan proporsi kelas yang seimbang, lanjutkan ke tahap analisis CART dalam proses analisis

A. Analisis CART tanpa Penerapan SMOTE

Langkah pertama dalam membuat pohon klasifikasi menggunakan teknik CART adalah memisahkan data ke dalam set pelatihan dan pengujian. Data *testing* digunakan untuk validasi model, dan data *training* digunakan untuk membuat pohon klasifikasi. Rasio data pelatihan dan pengujian dalam penelitian ini adalah 80% berbanding 20%, atau masing-masing 444 sampel untuk pelatihan dan 91 sampel untuk pengujian. Pohon keputusan adalah pohon kategorisasi. Pemisah utama, juga dikenal sebagai simpul akar yang adalah variabel dengan pengotor terbesar.

Memilih pemilah terbaik untuk dijadikan pemilah utama adalah langkah pertama dalam menggunakan metode CART. Pemilah utama yang diperoleh adalah jenis pencahayaan (X_{10}). Proses selanjutnya adalah mengurangi pohon klasifikasi maksimum menjadi pohon yang lebih sederhana yang dikenal sebagai pohon optimal. Lima simpul terminal dengan tiga variabel independen dengan jenis penerangan (X_{10}), sumber air minum (X_{11}), dan Fasilitas BAB (X_{13}) yang dihasilkan oleh pohon klasifikasi terbaik. Gambar 3 menunjukkan diagram teknik CART.



Gambar 3. Pohon Klasifikasi Optimal tanpa SMOTE

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/232

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Pada metode CART, ketepatan klasifikasi akan diukur melalui nilai akurasi, sensitivitas dan spesifisitas.

$$\text{Akurasi} = \frac{3+1}{112} = \frac{103}{112} = 0{,}92$$

$$\text{Sensitivitas} = \frac{3}{3+3} = \frac{1}{6} = 0{,}5$$

$$\text{Spesifisitas} = \frac{100}{100} = \frac{100}{106} = 0{,}94$$

Berdasarkan Tabel 3 ketepatan klasifikasi CART tanpa SMOTE menghasilkan nilai akurasi sebesar 92%, sebesar sensitivitas 50% dan spesifisitas sebesar 94%. Meskipun memiliki tingkat akurasi yang tinggi, model yang dibentuk masih belum baik dalam melakukan pengklasifikasian karena sensitivitas yang rendah sehingga mempengaruhi nilai *AUC* yang didapatkan.

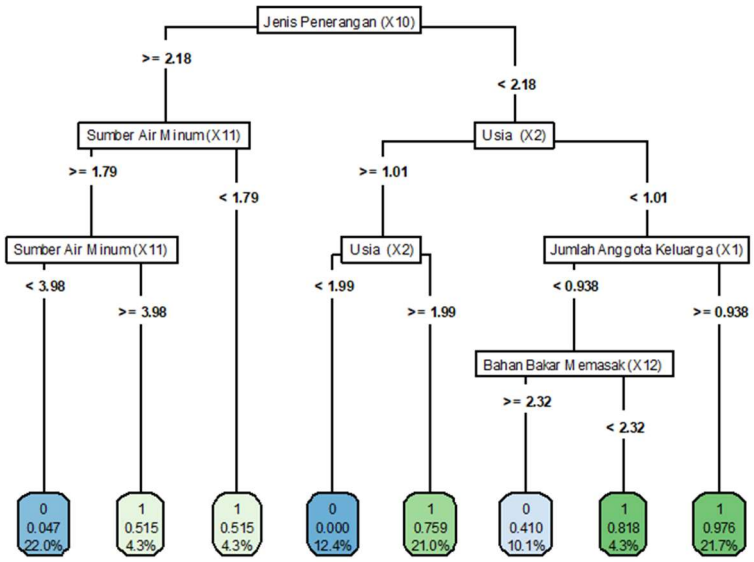
Tabel 3. Ketepatan Klasifikasi CART tanpa SMOTE

Aktual	Prediksi		Ketepatan
	Miskin	Tidak Miskin	
Miskin	3	3	50%
Tidak Miskin	6	100	94%
Akurasi Total			92%

B. Analisis CART dengan Penerapan SMOTE

Penerapan CART dengan SMOTE pada penelitian ini menggunakan data asli digabung dengan data sintesis hasil SMOTE. Sebelum membentuk model CART dengan SMOTE, membagi data ke dalam data *training* dan data *testing,* dimana 80% amatan digunakan sebagai data training dan 20% amatan digunakan sebagai data testing sehingga data training yang digunakan sebanyak 773 amatan dan terdapat 112 amatan pada data testing.

Tahap pertama dalam penerapan metode CART dengan SMOTE adalah menentukan pemilah terbaik sebagai *root node*. Variabel peubah jenis penerangan (X_{10}) merupakan pembagi yang paling efektif dalam analisis CART menggunakan SMOTE. Peubah jenis penerangan memiliki nilai impuritas yang paling tinggi dibandingkan dengan peubah independen lainnya. Maka diperoleh diagram metode CART dengan SMOTE dapat dilihat pada Gambar 4.



Gambar 4. Pohon Klasifikasi Optimal dengan SMOTE

Pohon klasifikasi optimal memiliki delapan simpul terminal dan mencakup enam variabel independen: sumber air minum (X_{11}), jumlah anggota keluarga (X_1), usia (X_2), bahan bakar memasak (X_{12}), dan jenis penerangan (X_{10}). Empat simpul terminal yang mengidentifikasi rumah tangga miskin sebagai rumah tangga miskin dan empat simpul terminal

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/232

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

yang mengidentifikasi rumah tangga miskin sebagai rumah tangga tidak miskin dihasilkan oleh pohon klasifikasi yang ideal. Mengukur ketepatan klasifikasi pohon keputusan dalam mengklasifikasikan kemiskinan rumah tangga dengan menggunakan metode CART dengan SMOTE:

$$\text{Akurasi} = \frac{5+90}{112} = \frac{95}{112} = 0{,}85$$

$$\text{Sensitivitas} = \frac{5}{5+1} = \frac{5}{6} = 0{,}83$$

$$\text{Spesifisitas} = \frac{90}{90+16} = \frac{90}{106} = 0{,}85$$

Tabel 4. Ketepatan Klasifikasi CART dengan SMOTE

Aktual	Prediksi		Ketepatan
	Miskin	Tidak Miskin	
Miskin	5	1	85%
Tidak Miskin	16	90	83%
Akurasi Total			85%

C. Perbandingan Tingkat Ketepatan Klasifikasi

Dengan membandingkan nilai *accuracy, sensitivity, specificity*, dan AUC dari dua pohon klasifikasi yang telah diperoleh sebelumnya, perbandingan dapat dilakukan. Tabel 5 menunjukkan hasil pengujian menggunakan data *testing* untuk membandingkan CART dengan dan tanpa SMOTE.

Tabel 5. Perbandingan Ketepatan Klasifikasi CART tanpa SMOTE dan dengan SMOTE

Kriteria	Tanpa SMOTE (%)	SMOTE (%)
Akurasi	92%	85%
Sensitivitas	50%	83%
Spesifisitas	94%	85%
AUC	72%	84%

Dari Tabel 5 terlihat bahwa secara keseluruhan pohon klasifikasi tanpa SMOTE memiliki akurasi yang lebih tinggi dibandingkan dengan pohon klasifikasi dengan SMOTE, namun demikian pohon klasifikasi tanpa SMOTE tidak dapat diklasifikasikan dengan baik untuk data rumah tangga kemiskinan. Hal ini ditunjukkan oleh nilai sensitivitas yang beda angka 50% atau bermakna sangat rendah. Kesalahan dalam klasifikasi kemiskinan rumah tangga yang miskin sebagai tidak miskin tentunya akan berakibat fatal. Setelah SMOTE selama tahap analisis pra-data, peningkatan sensitivitas hingga 83% diamati. Selain itu, nilai AUC dengan SMOTE terlihat lebih besar daripada nilai AUC tanpa SMOTE, yaitu masing-masing 72% dan juga 84%.

IV. KESIMPULAN

Kesalahan klasifikasi yang relatif besar pada kategori kelas minor disebabkan oleh data yang tidak seimbang. SMOTE digunakan untuk menyeimbangkan volume informasi pada setiap keluarga berpenghasilan rendah. Nilai akurasi dari teknik SMOTE adalah 85%, sedangkan pendekatan tanpa SMOTE memiliki nilai akurasi 92%. Namun terjadi peningkatan nilai sensitivitas sebesar 33%%. Sementara itu, teknik CART dengan SMOTE mengalami peningkatan nilai AUC sebesar 12%. Jadi pada penelitian ini dapat disimpulkan bahwa analisis klasifikasi CART dengan SMOTE dapat berkontribusi memberikan performa yang baik jika melihat kembali hasil analisis klasifikasi CART tanpa SMOTE. Sehingga metode CART dengan SMOTE merupakan metode terbaik dalam mengklasifikasikan kemiskinan rumah tangga di Kabupaten Kepulauan Mentawai, berdasarkan model CART dengan SMOTE. Dapat dilihat dari nilai *cunfunsion matriks* yang didapatkan bahwa menghasilkan ketepatan rumah tangga yang berstatus tidak miskin dengan sangat kecil. Saran untuk penelitian selanjutnya yaitu dapat mencobakan membandingkan dengan menggunakan metode seperti SVM, *Random Forest* dan lainnya

DAFTAR PUSTAKA

Badan Pusat Statistik (BPS). 2011. Kemiskinan dan Ketimpangan.

Badan Pusat Statistik (BPS). 2021. Kemiskinan dan Ketimpangan.

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/232

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Breiman, L., Friedman, J. H., Olshen, R. A., dan Stone, C. J. (1984) *Classification And Regression Trees*, New York: Chapman and Hall.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research Vol. 16, No. 2, Hal: 321–357.

Ispriyanti, D., Prahutama, A., Mustafid, M., & Tarno, T. KLASIFIKASI KEMISKINAN DI KOTA SEMARANG MENGGUNAKAN ALGORITMA CHISQUARE AUTOMATIC INTERACTION DETECTION (CHAID) DAN CLASSIFICATION AND REGRESSION TREE (CART). *Media Statistika*, *12*(1), 63-72.

Pratiwi, EF., Zain, I. *Klasifikasi Pengangguran Terbuka Menggunakan CART (Classification and Regression Tree).* Sulawesi Utara: *Jurnal sains pomits*, 3(1); 2014.

Sari, Veronica Retno, Feranandah Firdausi, & Yufis Azhar. (2020). Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naïve Bayes. Jurnal Pendidikan Informatika, 4(2), 1- 9.

Sartono B. Syafitri UD. 2010. Metode Pohon Gabungan: solusi pilihan untuk mengatasi kelemahan pohon regresi dan klasifikasi tunggal. From Statistika dan Komputasi. 15(1): 1-7.

Sumbar antarnews, "Fakta penyebab Kepulauan Mentawai masih tertinggal," sumbar antarnews, 2019. https://sumbar.antaranews.com/berita/281562/fakta-penyebab-kepulauan mentawai-masih-tertinggal (accessed Jun. 1, 2024).

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/232