ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 10, Nomor 2, Tahun 2021, Halaman 221-229
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

KLASIFIKASI STATUS KEMISKINAN RUMAH TANGGA DENGAN ALGORITMA C5.0 DI KABUPATEN PEMALANG

Fatiya Nur Umma[1], Budi Warsito[2], Di Asih I Maruddani[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
e-mail : fatiyaumma@gmail.com

ABSTRACT

Pemalang regency is a district which has amount of poverty around 16.04%. One of the effort that must be improved in tackling poverty is increasing the accuracy of the government program's target. The improvement of target accuracy is expected to give the better impact on the welfare of the population. This study classified the poverty status of households in Pemalang regency using C5.0 Algorithm. The poverty status of households is divided into two classes, namely poor and non-poor. There was an imbalance of data in both classes. Data imbalances were handled by using Synthetic Minority Oversampling Technique (SMOTE). From the research that has been done, SMOTE application in classification of household poverty status affected the evaluation value of the model. Previously the model could not classify the minority class and after using SMOTE the model produced an average value of sensitivity 25.80%. SMOTE application increased the average value of specificity from 91.16% to 94.91%. However, SMOTE application decreased the average value of accuracy which originally 91.16% down to 82.2%.

Keywords : C5.0, Household poverty, Classification, SMOTE

1. PENDAHULUAN

Kemiskinan menurut World Bank (2000) didefinisikan sebagai "*deprivation in well-being"* yang berarti kehilangan kesejahteraan. Kemiskinan merupakan sebuah kondisi dimana seorang atau sekelompok orang, laki-laki dan perempuan, tidak mampu memenuhi hak-hak dasarnya untuk mempertahankan dan mengembangkan kehidupan yang bermartabat. BPS (2004) menggunakan konsep kemampuan untuk memenuhi kebutuhan dasar sebagai alat untuk mengukur kemiskinan. Penduduk miskin diartikan sebagai penduduk yang memiliki rata-rata pengeluaran per kapita per bulan di bawah garis kemiskinan. Garis kemiskinan sendiri merupakan penjumlahan dari Garis Kemiskinan Makanan (GKM) yang merupakan nilai pengeluaran kebutuhan minimum makanan dan Garis Kemiskinan Non Makanan (GKNM) yang merupakan kebutuhan minimum untuk perumahan, sandang, pendidikan, dan kesehatan.

Pemalang merupakan kabupaten dengan dengan persentase kemiskinan sebesar 16,04%. Garis kemiskinan di Pemalang berada pada angka Rp 351.183 per bulan. Berbagai program telah dijalankan oleh pemerintah yang tujuannya untuk dapat mengurangi jumlah dan proporsi penduduk miskin. Sejumlah permasalahan masih ditemui meskipun pemerintah telah mengalokasikan dana yang sangat besar untuk penerapan kebijakan. Menurut Bappenas pemenuhan kebutuhan dasar atas masyarakat miskin atas pendidikan dan kesehatan menemui kendala pendataan dan akurasinya, sementara itu penyediaan prasarana dasar menjumpai masalah kelembagaan dan akurasi sasaran pemanfaat program. Peningkatan akurasi sistem pendataan agar program-program pemerintah semakin berdampak pada kesejahteraan penduduk miskin salah satunya dapat dilakukan dengan mengimplementasikan data mining.

Data mining akan diterapkan pada data kemiskinan sehingga dapat memberikan informasi tambahan dalam mengoptimalkan penanggulangan kemiskinan dengan menggunakan metode klasifikasi. Klasifikasi merupakan proses menemukan sebuah model atau fungsi yang mendeskripsikan dan mencirikan konsep atau kelas data untuk kepentingan tertentu. Metode yang ada dalam pengklasifikasian salah satunya adalah Pohon Keputusan.

Algoritma C5.0 merupakan salah satu Algoritma Pohon Keputusan dan merupakan pengembangan dari Algoritma C4.5. Model klasifikasi yang baik dapat dilihat dari tingkat akurasinya dalam memprediksi berdasarkan kategori respon. Hal yang dapat memengaruhi akurasi dari model klasifikasi salah satunya adalah masalah ketidakseimbangan data. Ketidakseimbangan data terjadi ketika salah satu kelas merepresentasikan jumlah data yang sangat besar, sedangkan kelas lainnya merepresentasikan data yang kecil. Pada kasus data tidak seimbang, sebagian besar Algoritma untuk klasifikasi cenderung mengklasifikasikan kelas mayoritas dengan tingkat akurasi tinggi dan kelas minoritas dengan tingkat akurasi rendah (Gu *et al.*, 2016). *Synthetic Minority Oversampling Technique* (SMOTE) merupakan satu dari beberapa alternatif untuk menyeimbangkan data pada kategori respon.

Pada penelitian ini mengkaji klasifikasi status kemiskinan rumah tangga di Kabupaten Pemalang dan penanganan data tidak seimbang dilakukan dengan membuat data sintesis baru pada data minoritas dengan metode SMOTE. Hasil akhirnya adalah mengetahui bagaimana pengaruh SMOTE yang digunakan untuk mengklasifikasikan status kemiskinan rumah tangga di Kabupaten Pemalang dengan Algoritma C5.0.

2. TINJAUAN PUSTAKA

2.1. Algoritma C5.0

Algoritma C5.0 adalah salah satu algoritma klasifikasi dalam data mining yang khususnya diterapkan pada Pohon Keputusan. C5.0 merupakan penyempurnaan algoritma sebelumya yaitu ID3 dan C4.5 yang dibentuk oleh Ross Quinland tahun 1987. Pemilihan atribut dalam C5.0 menggunakan nilai *gain ratio*. *Gain ratio* berdasar pada konsep *Entropy* dalam perhitungannya. *Entropy* adalah suatu parameter untuk mengukur tingkat keberagaman atau heterogenitas dari kumpulan data. Rumus yang digunakan untuk menghitung *Entropy* adalah:

$$Entropy\ (S) = -\sum_{i=1}^{n} P_i * log_2 P_i \quad (1)$$

dengan S = himpunan data

n = jumlah kelas pada variabel target

Pi = proporsi banyaknya data kelas ke-i pada himpunan data

Selanjutnya akan dihitung nilai *information gain* variabel prediktor. *Information gain* adalah ukuran efektifitas suatu variabel prediktor dalam mengklasifikasikan data. Perhitungan gain disajikan sebagai berikut.

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} * Entropy(Si) \quad (2)$$

dengan A = variabel prediktor

k = banyaknya kategori pada variabel prediktor A

$|Si|$ = jumlah sampel untuk kategori ke-i

$|S|$ = jumlah seluruh sampel dalam himpunan data

Perhitungan *gain ratio* untuk Algoritma C5.0 akan berjalan setelah *information gain* dilakukan. Nilai *gain ratio* dihitung menggunakan rumus berikut.

$$Gain\ Ratio = \frac{Gain\ (S,A)}{SplitInfo\ (S,A)} \quad (3)$$

dengan $Gain(S, A)$ = *information gain* pada variabel prediktor A

$SplitInfo(S,A)$ = *split information* pada variabel prediktor A

Nilai *gain ratio* tertinggi dipilih sebagai atribut untuk simpul. Pendekatan ini menerapkan normalisasi pada *information gain* dengan menggunakan apa yang disebut sebagai *split information*. SplitInfo menyatakan informasi potensial dengan persamaan 4.

$$SplitInfo(S,A) = -\sum_{i=1}^{k} \frac{S_i}{S} log_2 \frac{S_i}{S} \quad (4)$$

2.2. *Synthetic Minority Oversampling Technique* (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) merupakan metode oversampling dengan membuat data sintetis atau data buatan secara acak (Chawla *et al.*, 2002). Metode SMOTE bekerja dengan mencari k-tetangga terdekat untuk setiap data di kategori minor, setelah itu dibentuk data sintesis sebanyak persentase duplikasi yang diinginkan antara data minoritas dan k-tetangga terdekat yang dipilih secara acak. Perhitungan jarak antara data yang akan dibangkitkan dengan tetangga terdekatnya menggunakan jarak Euclidean. Ketika data bertipe numerik dan kategorik, perhitungan jarak tetap menggunakan jarak Euclidean namun menggunakan nilai median dari simpangan baku peubah numerik sebagai selisih nilai peubah kategorik. Nilai median ini dihitung ketika nilai kategori amatan dan tetangga terdekatnya berbeda. Rumus perhitungan jarak Euclidean yaitu:

$$(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (5)$$

Pembangkitan data buatan yang berskala numerik berbeda dengan kategorik, pembangkitan data buatan dilakukan dengan cara sebagai berikut:

a. Data Numerik

Data yang memiliki jarak terdekat kemudian digunakan untuk membangkitkan data sintesis baru berdasarkan persamaan 6.

$$X_{baru} = x + (x^* - x) \times rand[0,1] \quad (6)$$

dengan X_{baru} = data sintesis hasil dari replikasi
x = data yang akan direplikasi
x^* = data yang memiliki jarak terdekat dari data yang akan di replikasi
Rand[0,1] = bilangan acak antara 0 sampai 1

b. Data kategorik

Pada data kategorik, data buatan dibangkitkan dengan menentukan kategori yang paling sering muncul diantara amatan dengan k-tetangga terdekat (modus). Apabila terdapat kesamaan nilai dari kategorinya, maka akan dipilih secara acak.

2.3. Evaluasi Model

Evaluasi model perlu dilakukan untuk mengetahui seberapa baik kinerja model klasifikasi yang terbentuk. Evaluasi model yang digunakan pada penelitian ini berupa matriks konfusi, yaitu matriks tabulasi silang antara kategori kelas sebenarnya dengan kelas hasil prediksi. Performa dari setiap model klasifikasi dapat dievaluasi dengan menggunakan 3 perhitungan statistik, yaitu klasifikasi keakuratan, sensitivitas, dan spesifisitas. Ketiganya ditentukan oleh *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN).

Tabel 1. *Confusion Matrix*

	Positif (Aktual)	Negatif (Aktual)
Positif (Prediksi)	*True Positive* (TP)	*False Positive* (FP)
Negatif (Prediksi)	*False Negative* (FN)	*True Negative* (TN)

Perhitungan akurasi, sensitivitas, dan spesifisitas secara berturut-turut menggunakan rumus sebagai berikut:

$$Akurasi = \frac{banyaknya\ prediksi\ yang\ benar}{total\ banyaknya\ prediksi} = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

$$Sensitivitas = \frac{banyaknya\ true\ positif}{banyaknya\ aktual\ positif} = \frac{TP}{TP+FN} \quad (8)$$

$$Spesifisitas = \frac{banyaknya\ true\ negatif}{banyaknya\ aktual\ negatif} = \frac{TN}{TN+FP} \quad (9)$$

3. METODE PENELITIAN

Jenis data yang digunakan dalam penelitian ini berupa data sekunder yang diperoleh dari Badan Pusat Statistik (BPS) Provinsi Jawa Tengah. Data merupakan data dari hasil Survei Ekonomi Nasional (SUSENAS) tahun 2018 dengan unit observasi adalah unit rumah tangga di Kabupaten Pemalang. Adapun variabel yang digunakan adalah sebanyak 19 variabel dimana terdiri dari variabel target (1=miskin; 0=tidak miskin) dan 18 variabel prediktor. Variabel prediktor yang digunakan diantaranya X1: jenis kelamin kepala RT (1=Laki-Laki; 2=perempuan); X2: usia kepala RT; X3: ijazah tertinggi kepala RT (1=Tidak memiliki ijazah; 2=SD sederajat; 3=SMP sederajat; 4=SMA sederajat; 5=D1/D2/D3; 6=D4/S1; 7=S2/S3); X4: banyaknya anggota RT; X5: status kerja kepala RT (1=Ya; 2=Tidak); X6: lapangan usaha (0=Tidak bekerja; 1=Pertanian tanaman padi dan palawija; 2=Hortikultura; 3=Perkebunan; 4=Perikanan; 5=Peternakan; 6=Kehutanan dan pertanian lainnya; 7=Pertambangan dan penggalian; 8=Industri pengolahan; 9=Pengadaan listrik, gas, uap/air panas, dan udara dingin; 10=Pengelolaan air, air limbah, daur ulang sampah, dan aktivitas remediasi; 11=Konstruksi; 12=Perdagangan besar dan eceran, reparasi dan perawatan mobil dan sepeda motor; 13=Pengangkutan dan pergudangan; 14=Penyediaan akomodasi dan makan minum; 15=Aktivitas keuangan dan asuransi; 16=Real estate; 17=Aktivitas penyewaan dan sewa guna tanpa hak opsi, ketenagakerjaan, agen perjalanan, dan penunjang usaha lainnya; 18=Administrasi pemerintahan, pertahanan, dan jaminan sosial wajib; 19=Pendidikan; 20=Aktivitas kesehatan manusia dan sosial; 21=Kesenian, liburan dan rekreasi; 22=Aktivitas jasa lainnya); X7: status dalam pekerjaan utama (0=Tidak bekerja; 1=berusaha sendiri; 2=berusaha dibantu buruh tidak tetap; 3=berusaha dibantu buruh tetap; 4=buruh/karyawan/pegawai; 5=pekerja bebas; 6=pekerja keluarga); X8: status kepemilikan tempat tinggal (1=milik sendiri; 2=Kontrak/Sewa; 3=Bebas sewa; 4=Dinas); X9: kemampuan membaca huruf latin (1=Ya; 2=Tidak); X10: penerima rastra (1=Ya; 2=Tidak); X11: bahan utama dinding rumah (1= Anyaman bambu; 2=Bambu; 3=Batang kayu; 4=Kayu/papan; 5=Plesteran anyaman bambu/kawat; 6=Tembok; 7=Lainnya); X12: bahan bangunan atap (1=Beton; 2=Asbes; 3=Bambu; 4=Genteng; 5=Jerami; 6=Seng; 7=Lainnya); X13: bahan utama lantai rumah (1=Marmer/granit; 2=Bambu; 3=Keramik; 4=Perket/vini/karpet; 5=Ubin/tegel/teraso; 6=semen/bata merah; 7=Lainnya); X14: kepemilikan fasilitas tempat buang air besar (1=Ada digunakan hanya ART sendiri; 2=Ada, digunakan bersama dengan ART lainnya; 3=Ada, ART tidak menggunakan; 4=Ada, di MCK Umum; 5= Tidak ada); X15: sumber utama air minum (1=Air kemasan bermerk; 2=Air isi

ulang; 3= Air hujan; 4=Air permukaan;5=Leding; 6=Sumur bor/pompa; 7=Sumur terlindung; 8=Sumur tak terlindung; 9=Mata air terlindung; 10=Mata air tidak terlindung); X16: tempat pembuangan akhir tinja (1=Tangki septik; 2= Pantai/tanah lapang/kebun; 3=IPAL; 4=Kolam/sawah/sungai/danau; 5=Lubang tanah; 6=Lainnya; 7=Tidak Memiliki); X17: Sumber utama penerangan (1=Listrik PLN dengan meteran; 2=Listrik meteran tanpa meteran; 3=Listrik non PLN; 4=Bukan listrik); X18: bahan bakar utama memasak (0=Tidak memasak; 1=Elpiji 3kg; 2=Ellpiji 5,5kg; 3=Elpiji 12kg; 4=Minyak tanah; 5=kayu bakar; 6=Lainnya).

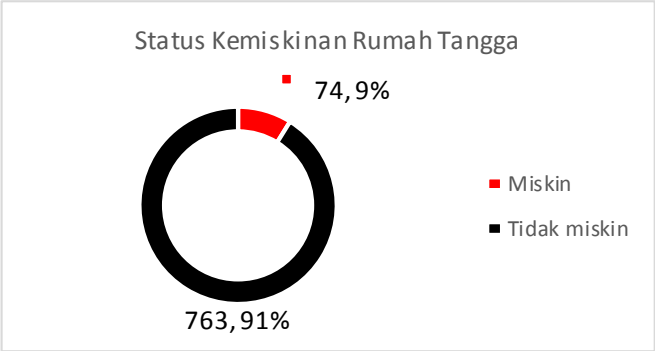
Tahap-tahap analisis data yang dilaksanakan sebagai berikut:

1. Mengklasifikasi status kemiskinan rumah tangga sesuai dengan definisi BPS dengan melihat pengeluaran perkapita per bulan dan dibandingkan dengan garis kemiskinan yang berlaku di Kabupaten Pemalang.
2. Melakukan eksplorasi data dan analisa statistika deskriptif untuk melihat karakteristik umum dari variabel-variabel yang akan dianalisis.
3. Melakukan *splitting* data yang dilakukan secara acak menjadi data latih dan data uji dengan *10-fold Cross Validation*.
4. Proses klasifikasi dengan Algoritma C5.0 akan dilakukan sebanyak k=10 kali dengan menggunakan data partisi ke-k sebagai data uji dan sisanya sebagai data latih.
5. Mengkur *performance* klasifikasi yaitu dengan rata-rata dari seluruh proses pelatihan.
6. Penanganan ketidakseimbangan data menggunakan metode SMOTE dengan langkah sebagai berikut:
   a. Menentukan jumlah k tetangga terdekat yaitu k=5, yang artinya data bangkitan berasal dari 5 data pada kelas minor yang berdekatan.
   b. Menentukan persentase *oversampling* sebesar 900% yang artinya data pada data sintesis yang akan dibuat yaitu sebanyak 9 kali dari kelas minor dan sebesar 150% pada kelas mayor.
   c. Mengulang langkah a dan b pada masing-masing iterasi hasil *splitting* data.
   d. Menggunakan data keluaran hasil SMOTE sebagai data latih baru.
7. Melakukan klasifikasi dengan Algoritma C5.0 dengan data latih baru hasil dari proses SMOTE sebanyak k=10 kali dengan menggunakan data partisi ke-k sebagai data uji dan sisanya sebagai data latih.
8. Mengevaluasi *performance* klasifikasi dengan rata-rata dari seluruh proses pelatihan model klasifikasi yang telah melalui SMOTE.
9. Membandingkan hasil evaluasi model untuk data yang melalui proses SMOTE dan tanpa melalui proses SMOTE untuk melihat pengaruh dari kinerja SMOTE terhadap data tidak seimbang.

4. HASIL DAN PEMBAHASAN

4.1. Analisis Deskriptif

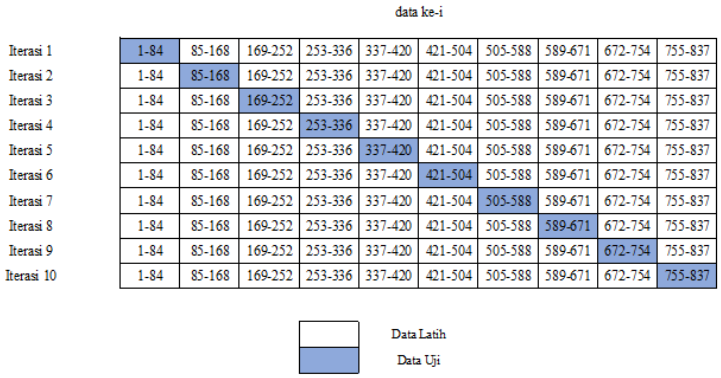
Data yang digunakan pada penelitian sebanyak 19 variabel. Variabel y merupakan variabel target dengan 2 kategori, yaitu miskin dan tidak miskin. Variabel x merupakan variabel prediktor yang terdiri dari 18 variabel, dimana terdapat 2 variabel pediktor yang bertipe numerik yaitu variabel usia dan jumlah anggota keluarga dan sisanya bertipe data kategorik. Eksplorasi data dilakukan untuk melihat karakteristik data dari setiap kategori. Telah disebutkan sebelumnya bahwa terdapat ketidakseimbangan data yang terjadi pada variabel target, yaitu variabel status kemiskinan. Proporsi ketidakseimbangan data dari kedua kategori ditunjukkan pada Gambar 1 berikut.



Gambar 1. Diagram Persentase Variabel

4.2. Pembagian Data

Pembagian data latih dan data uji pada penelitian ini menggunakan *k-fold cross validation* dengan k=10. Jumlah data yang digunakan adalah sebanyak 837 observasi dengan pembagian data latih dan data uji diilustrasikan seperti pada Gambar 2.



Gambar 2. Ilustrasi Pembagian Data Latih dan Data Uji

Tabel 2. Evaluasi Hasil Klasifikasi dengan Algoritma C5.0

	Akurasi	Sensitivitas	Spesifisitas
Fold 1	90,48%	Na	90,48%
Fold 2	95,24%	Na	95,24%
Fold 3	90,48%	Na	90,48%
Fold 4	88,10%	Na	88,10%
Fold 5	90,48%	Na	90,48%
Fold 6	92,86%	Na	92,86%
Fold 7	91,67%	Na	91,67%
Fold 8	90,36%	Na	90,36%
Fold 9	90,36%	Na	90,36%
Fold 10	91,57%	Na	91,57%

4.3. Permodelan Klasifikasi Status Rumah Tangga dengan Algoritma C5.0

Model yang terbentuk dari masing-masing *fold* selanjutnya akan diuji kinerjanya menggunakan data uji. Penelitian ini menggunakan *confusion matrix* dalam proses evaluasi untuk mengetahui nilai dari akurasi, sensitivitas, maupun spesifisitas. Proses klasifikasi C5.0 dengan 10-*fold cross validation* menghasilkan tingkat akurasi, sensitivitas, dan spesifisitas yang berbeda dari masing-masing nilai k yang digunakan. Hasil dari masing-masing *fold* disajikan pada Tabel 2.

4.4 Permodelan Klasifikasi C5.0 Dengan Menggunakan SMOTE

Masalah ketidakseimbangan data ditangani menggunakan metode SMOTE yang diterapkan terhadap data latih. Teknik SMOTE akan digunakan untuk menangani ketidakseimbangan data. Teknik ini membangkitkan data sintesis untuk kelas minoritas sebesar 900% atau sebanyak 9 kali dari data latih minoritas dan 150% dari data mayoritas. Data baru yang dihasilkan nantinya akan ditambahkan ke data awal. Tabel 3 memuat salah satu contoh perbandingan data latih awal dengan data latih yang telah melalui proses SMOTE.

Tabel 3. Persentase Data Sebelum dan Sesudah SMOTE

Kategori	Banyaknya Data Awal (%)	Banyaknya Data setelah SMOTE (%)
1= miskin	66 (10%)	660 (42,55%)
0= tidak miskin	594 (90%)	891 (57,45%)
Jumlah	660 (100%)	1551 (100%)

Setelah dilakukan penyeimbangan data dengan metode SMOTE, data latih baru digunakan untuk membangun model *classifier* yang baru dengan metode yang sama pula, yaitu dengan 10-*fold cross validation* dan Algoritma C5.0. Evaluasi model dilakukan pada data uji dan menghasikkan kinerja klasifikasi seperti pada Tabel 4 berikut.

Tabel 4. Evaluasi Hasil Klasifikasi dengan Algoritma C5.0 dengan SMOTE

	Akurasi	Sensitivitas	Spesifisitas
Fold 1	80,95%	25,00%	94,12%
Fold 2	84,52%	15,38%	97,18%
Fold 3	79,76%	15,38%	91,55%
Fold 4	77,38%	30,43%	95,08%
Fold 5	82,14%	26,67%	94,20%
Fold 6	82,14%	15,38%	94,37%
Fold 7	83,33%	29,41%	97,02%
Fold 8	81,93%	31,58%	96,87%
Fold 9	90,36%	50,00%	94,67%
Fold 10	79,52%	18,75%	94,03%

Hasil klasifikasi dengan algoritma C5.0 menghasilkan ukuran tingkat kepentingan pada masing-masing variabel prediktor yang digunakan. Tingkat kepentingan tersebut

menggambarkan seberapa besar kontribusi dari variabel prediktor dalam membagun suatu model klasifikasi. Secara keseluruhan, untuk model klasifikasi status rumah tangga miskin dengan algoritma C5.0 yang sebelumnya telah menggunakan metode SMOTE menghasilkan pohon klasifikasi dengan peubah penting yang berbeda-beda. Masing-masing *fold* akan menghasilkan tingkatan kepentingan variabel prediktor dengan urutan dan persentase yang berbeda-beda. Variabel prediktor yang paling sering digunakan pada keseluruhan *fold* dalam pembentukan pohon keputusan adalah Bahan Bangunan Utama Dinding, Banyaknya Anggota RT, Sumber Utama Air Minum, Bahan Bakar Utama Memasak, dan variabel Kemampuan Membaca Huruf Latin.

4.5 Perbandingan Model Klasifikasi

Perbandingan hasil evaluasi model dengan SMOTE dan tanpa SMOTE dilakukan dengan membandingkan nilai hasil dari rata-rata dari akurasi, sensitivitas, dan spesifisitas. Nilai rataan akurasi, sensitivitas, dan spesifisitas pada masing-masing model dapat dilihat pada Tabel 5.

Tabel 5. Perbandingan Hasil Model Klasfikasi tanpa SMOTE dan dengan SMOTE

Kriteria	Model	
	Tanpa SMOTE	Dengan SMOTE
Akurasi	91,16%	82,20%
Sensitivitas	Na	25,80%
Spesifisitas	91,16%	94,91%

Tabel 5 menunjukkan rataan *performance* dari kedua model yang merupakan rata-rata dari 10 percobaan *machine learning* yang telah dilakukan. Model klasifikasi data sebelum menggunakan SMOTE menghasilkan rata-rata akurasi sebesar 91,16%, sedangkan setelah dilakukan penanganan dengan menggunakan SMOTE rataan akurasi turun menjadi 82,20%. Penggunaan SMOTE untuk menangani ketidakeimbangan data menaikkan nilai spesifisitas yang sebelumnya 91,16% menjadi 94,91%. Rentannya pohon klasifikasi terhadap ketidakseimbangan data dapat jelas terlihat pada nilai sensitivitas yang dihasilkan. Sebelum menggunakan SMOTE, nilai dari sensitivitas yang merupakan derajat keberhasilan sebuah model untuk mengklasifikasi data dengan kategori positif, dalam hal ini merupakan data rumah tangga dengan kategori miskin tidak dapat didefinisikan. Hal tersebut menunjukkan tidak berhasilnya C5.0 dalam memprediksi kelas minoritas. Penggunaan SMOTE untuk menangani data tidak seimbang memberikan perbedaan hasil klasifikasi dibandingkan dengan data sebelum melalui proses SMOTE. Dalam Tabel 5, setelah dilakukan penanganan data tidak seimbang menggunakan SMOTE, model dapat mengklasifikasi kelas dari kategori minor sehingga nilai rata-rata sensitivitas menjadi sebesar 25,80%.

Berdasarkan perbandingan rataan *performance* dari kedua model tersebut, dapat dilihat bahwa metode SMOTE mampu memperbaiki hasil klasifikasi dengan Algoritma C5.0. Model tanpa SMOTE menunjukkan angka yang tinggi pada perhitungan akurasi dan spesifisitas, akan tetapi model tersebut tidak dapat digunakan karena tidak berhasil mengklasifikasi data dari kelas minor. Hal tersebut dikarenakan perbedaan jumlah amatan yang terlalu tinggi sehingga *machine learning* hanya fokus untuk mengklasifikasi data dalam kelas mayoritas. Penanganan data dengan SMOTE meningkatkan kemampuan *machine learning* sehingga dapat mengklasifikasi data dalam kelas minor. SMOTE berhasil menghasilkan rataan sensitivitas menjadi sebesar 25,80%.

5. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan yang dilakukan pemodelan klasifikasi status kemiskinan rumah tangga di Kabupaten Pemalang menggunakan Algoritma C5.0 menghasilkan nilai rata-rata akurasi sebesar 91,16% dan rata-rata spesifisitas sebesar 91,16%. Akan tetapi, nilai sensitivitas tidak dapat terdefinisi sehingga model dianggap tidak berhasil dalam mengklasifikasi data dengan kategori sebagai rumah tangga miskin. Klasifikasi status kemiskinan rumah tangga di Kabupaten Pemalang menggunakan Algoritma C5.0 yang telah melalui proses *Synthetic Minority Oversampling Technique* (SMOTE) menghasilkan rataan nilai akurasi, sensitivitas, dan spesifisitas secara berturut-turut sebesar 82,20%, 25,80%, dan 94,91%. Penggunaan SMOTE sebagai metode *pre-processing* data meningkatkan kemampuan dari machine learning sehingga dapat mengklasifikasi data baik dari kelas mayor maupun minor meskipun belum menunjukkan hasil yang cukup signifikan.

DAFTAR PUSTAKA

Arif, M. 2018. Decision Tree Algorithms C4.5 and C5.0 in Data Mining: A Review. *International Journal of Databases Theory and Application*. Vol.11, No.1, hal.1-8.

Blagus, R dan Lara, L. 2013. SMOTE for High-Dimensional Class-Imbalanced Data. *BMC Bioinformatics*, Vol. 14, hal.106.

Chawla, N. V., Bowyer, K. W., Hall, L. O., dan Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, Vol. 16, hal. 321–357.

Han, J dan Kamber, M. 2006. *Data mining: concepts and techniques*. 2nd edition. San Francisco: Morgan Kauffman.

Kantardzic, M. 2011, *Data Mining: Concept, Models, Methods, and Algorithms*. 2nd edition. New Jersey: John W & Sons, Inc.

Larose, D. T, 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: Jhon Wiley & Sons Inc.

Qiong, Gu., Wang, X.-M., Zhao, Wu., Ning, B., dan Xin, C.-S.,. 2016. An Improved SMOTE Algorithm Based On Genetic Algorithm For Imbalanced Data Classification. *Journal of Digital Information Management*, Vol. 14, No. 2, hal. 92–103.

Susanto, Sani dan Dedy Suryadi. 2010. *Pengantar Data Mining Menggali Pengetahuan dari Bongkahan Data*. Yogyakarta: Andi.

Tan, P., Steinbach, M., dan Kumar, V. 2006. *Introduction to Data Mining*. Boston: Pearson Education.

Witten, I. H dan Frank, E. 2005. *Data Mining Practical Machine Learning Tools and Techniques*. 2nd Edition. San Fransisco: Morgan Kaufmann.