KLASIFIKASI KARAKTERISTIK KEMISKINAN DI PROVINSI BENGKULU TAHUN 2020 MENGGUNAKAN METODE POHON KLASIFIKASI GABUNGAN

Winalia Agwil[1], Dian Agustina[2], Herlin Fransiska[3], Nurul Hidayati[4]

[1,2,3,4] Program Studi Statistika, FMIPA Universitas Bengkulu, Bengkulu
e-mail: [1]winaliaagwil@unib.ac.id

Abstrak

Kemiskinan adalahmasalah yang mendesak diatasi baik pada tingkat nasional maupun global, yang diindikasikan sebagai salah satu prioritas utama dalam agenda dunia tentang Tujuan-tujuan Pembangunan Berkelanjutan atau *Sustainable* Development *Goals* (SDGs). Penanganan kemiskinan yang efektif akan membantu penyelesaian permasalahan dunia yang lain seperti permasalahan kelaparan, kesehatan, kesejahteraan, pendidikan, air bersih dan sanitasi. Studi ini bertujuan untuk mengklasifikasikan karakteristik kemiskinan rumahtangga di Propinsi Bengkulu, berdasarkan data hasil Survei Sosial Ekonomi Nasional (SUSENAS) 2020. Metode analisis data untuk mengidentifikasi karakteristik rumah tangga miskin dalam studi ini menggunakan *Classification and Regression Tree* (CART) – dengan menerapkan model random forest untuk mengatasi ketidakseimbangan data. Berdasarkan pemodelan tersebut, studi ini menemukan bahwa terdapat tiga variabel utama yang mencirikan rumah tangga miskin di propinsi Bengkulu, yaitu jumlah anggota rumah tangga, ijazah terakhir kepala rumah tangga dan luas lantai rumah. Temuan ini dapat digunakan sebagai dasar untuk identifikasi rumahtangga miskin, sehingga program-program bantuan diharapkan lebih tepat sasaran di masa mendatang.

Kata Kunci: Kemiskinan, Unbalanced dataset, CART, Random forest, Xgboost

Abstract

Poverty is a pressing issue at both country and global level, indicated as one of main priorities in the global agenda on Sustainable Development Goals (SDGs). Tackling poverty effectively will help solve other critical issues such as hunger, health, welfare, education, clean water, and sanitation. The aim of this study is to classify the main characteristics of poor households in Province of Bengkulu, using the data from the results of National Socioeconomic Survey (SUSENAS) 2020. Method of data analysis in identifying the poor household characteristics in the present study uses Classification and Regression Tree (CART) – specifically by applying a forest random model in attempt to adjust the unbalanced datasets. Main findings from this modelling application suggest that there are three variables that mainly characterized poor households in Bengkulu, namely the number of household members, the last education certificate of the household head, and the size of the house floor area. This finding is useful for a basis in identifying the poor households, thus intervention programs designed to assist the poor is expected to be more well-targeted in the future.

Keywords*: Poverty, Unbalance datasets, CART, Random forest*

PENDAHULUAN

Kemiskinan merupakan isu mendesak untuk diatasi di tingkat nasional dan deklarasi *Millenium Development Goals* (MDGs) yang salah satu targetnya adalah mengurangi penduduk miskin dunia. Era MDGs berhasil mencapai tujuannya yakni mengatasi kemiskinan hingga 50%. Selanjutnya memasuki era *Sustainable Development Goals* (SDGs) yang dicetus untuk meneruskan dan memantapkan capaian MDGs agar berlanjut. SDGs memprioritaskan "*no poverty*", karena mengatasi kemiskinan secara efektif akan sejalan dengan upaya-upaya mengatasi permasalahan dunia lainnya seperti: dunia tanpa kelaparan, kesehatan yang baik, kesejahteraan, pendidikan berkualitas, air bersih dan sanitasi (Ishartono dan Raharjo, 2016). Pada tahun 2020, Provinsi Bengkulu merupakan provinsi dengan persentase penduduk miskin kedua tertinggi di pulau Sumatera setelah Aceh, yaitu mencapai 14,43%. Mengetahui karakteristik rumah tangga miskin di Provinsi Bengkulu menjadi penting untuk dikaji, sebagai acuan dalam penyaluran program pengetasan kemiskinan sehingga program pemerintah dapat dilakukan secara tepat sasaran dan efisien.

Pengkajian tentang kemiskinan dapat dilihat dari unit paling kecil yakni rumah tangga. Menurut BPS, rumah tangga dikategorikan sebagai miskin, jika nilai rata-rata pengeluaran konsumsi per kapita per bulan dibawah garis kemiskinan.Dan sebaliknya, jika nilai rata-rata pengeluaran konsumsi per kapita per bulan berada diatas garis kemiskinan maka rumah tangga tersebut dapat di kategorikan tidak miskin. Penentuan karakteristik dari rumah tangga miskin dan tidak miskin dapat dilakukan dengan analisis klasifikasi. Metode klasifikasi yang paling umum digunakan antara lain, regresi logistik, klasifikasi pohon, K-tetangga terdekat, jaringan saraf tiruan, dan Support Vector Machine (SVM). Setiap metode memiliki syarat/batasan untuk dapat digunakan juga memiliki kelebihan dan kekurangan. Secara umum, permasalahan ketidakseimbangan kelas data menyebabkan analisis klasifikasi pohon tunggal akan menghasilkan pohon yang kurang stabil dimana ketika data training mengalami perubahan kecil dapat memberikan perubahan yang signifikan pada pohon yang dihasilkan (Sutton, 2005).

Ispriyanti, Prahutama dan Mustafid (2019) juga eneliti pengklasifikasian kemiskinan di Kota Semarang dengan menerapkan metode algoritma QUEST. Metode QUEST termasuk salah satu dari banyak metode pohon klasifikasi tunggal. Penelitian tersebut menunjukkan perbandingan antara jumlah rumah tangga miskin dan tidak miskin sebesar 47:883. Hal ini mengindikasikan adanya ketidakseimbangan kelas data yang potensial mengakibatkan kesalahan klasifikasi saat memprediksi kelas data. Hal ini misalnya hasil klasifikasi pada awalnyadiperoleh klasifikasi rumah tangga miskin, namun setelah di prediksi klasifikasi berubah menjadi rumah tangga tidak miskin.

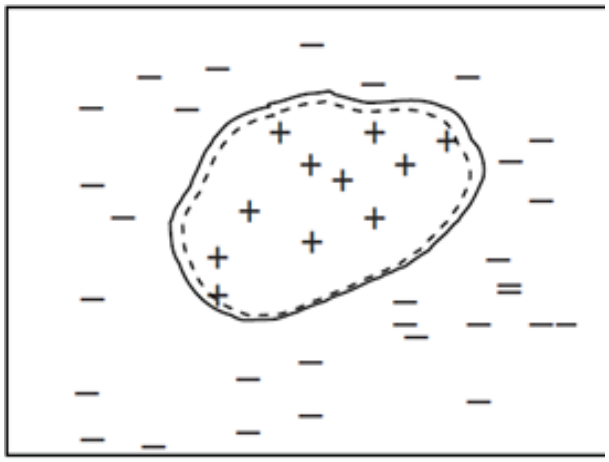
Kondisi ketidakseimbangan data perlu untuk ditangani sehingga memperbaiki performa model klasifikasi, penulis tertarik untuk menangani permasalahan ini dengan menerapkan metode Synthetic Minority Over-sampling Technique (SMOTE) yang merupakan solusi pada level data dan juga menerapkan metode pohon klasifikasi gabungan seperti Random Forest (RF) dan Extreme Gradien Boosting (Xgboost) untuk pemodelan klasifikasi.

Tujuan penelitian ini adalah untuk mengidentifikasi karakteristik rumah tangga miskin dan tidak miskin di Provinsi Bengkulu, kemudian melakukan pengklasifikasian status kemiskinan rumah tangga dengan menggunakan metode CART, Random Forest dan Extreme Gradient Boosting (Xgboost). Selain itu, masalah ketidakseimbangan data ditanggulangi dengan metode SMOTE yang kemudian datanya akan digunakan untuk pengklasifikasian rumah tangga dengan menggunakan metode CART, Random Forest Extreme Gradient Boosting (Xgboost). Dimana output dari penelitian adalah menghasilkan model pohon klasifikasi yang menggambarkan karakteristik kemiskinan di Provinsi Bengkulu. Model dibentuk dengan menggunakan data hasil SUSENAS Maret 2020 modul KOR dan Modul Konsumsi.

METODOLOGI

Ketidakseimbangan Data

Ketidakseimbangan data terjadi apabila pada suatu dataset terdapat suatu kategori (kelas) data yang mendominasi kategori yang lainnya. Kategoridata yang memiliki jumlah observasi yang lebih banyak disebut dengan kelas mayoritas (kelas negatif), sedangkan kategori data yang memiliki jumlah observasi lebih sedikit disebut dengan kelas minoritas (kelas positif) (Chawla et al. 2002). Ketidakseimbangan data diilustrasikan pada Gambar 1.



Gambar 1 Data Tidak Seimbang
Sumber: Chawla et al. 2002

Pada Gambar 1, dapat dilihat tanda "+" menggambarkan obsevasi yang berasal dari kategori minoritas dan "-" mewakili observasi dari kelas mayoritas.

Penanganan ketidakseimbangan data dapat dilakukan dengan menggunakan Synthetic Minority Oversampling Technique (SMOTE). Metode SMOTE pertama kali dikenalkan oleh Chawla et.al (2002) yang merupakan metode penanggulangan ketidakseimbangan data dengan melakukan resampling pada data.

Metode SMOTE adalah proses penambahan data baru pada kelas minoritas agar jumlah observasinya sebanding dengan kelas mayoritas, yang dilakukan berdasarkan informasi tetangga terdekat (k-nearest neighbor). Terdapat dua tahapan dalam metode ini, yang pertama tahapan menentukan tetangga terdekat untuk setiap observasi yang berada pada kelas minoritas dengan menggunakan jarak Euclidean jika data berupa numerik dan Value Difference Metric (VDM) jika data kategorik. Tahapan kedua adalah pembuatan data sintetik. Pada data numerik, data baru dihitung dengan persamaan $x^* = x_i + (bilangan\ acak(0 - 1)\ selisih(x_i, x_{ij})$. Sedangkan untuk data kategorik, observasi baru merupakan mayoritas nilai dari k-tetangga terdekatnya (Chawla et al. 2002).

Classification and Regression Tree (CART)

Metode CART merupakan salah satu pohon keputusan popular yang dapat digunakan pada variabel respon numerik maupun kategorik. Jika variabel respon berupa data kategorik dinamakan pohon klasifisasi, sedangkan variabel respon yang berupa data numerik dinamakan pohon regresi. Pembentukan pohon klasifikasi maupun pohon regresi dilakukan dengan proses rekursif biner pada gugus data sehingga pada pemilahan terakhir diperoleh nilai variabel respon pada setiap simpul yang terbentuk lebih homogen (Breiman et al. 1984).

Lewis (2000) menyebutkan metode ini sebagai klasifikasi *binary recursive partitioning*, karena setiap simpul yang dihasilkan disekat atau dipisahkan menjadi dua simpul anak. Tahapan penyekatan tersebut dilakukan sampai terpenuhi kriteria pemberhentian yang ditetapkan. Algoritma CART secara umum digambarkan dalam tahapan berikut (Breiman et. Al. 1984):

1. Menemukan pemisah terbaik pada setiap variabel prediktor yang digunakan dalam model. Setiap variabel dengan nilai K yang berbeda memiliki k-1 dengan kemungkinan pemisah. Pemisah terbaik dipilih berdasarkan kriteria *splitting* yaitu *Gini's impurity*. Setiap variabel prediktor memiliki satu nilai pemisah terbaik. Berikut adalah formula indeks Gini:

$$i(t) = -\sum_{j=1}^{J} p(j|t) log_2 p(j|t) \qquad (1)$$

$p(j|t)$ merupakan peluang observasi kelas j pada *simpul* t. Evaluasi pemisah s pada simpul t dapat dilakukan dengan melihat nilai *Goodness of split*.

2. Menemukan variabel prediktor terbaik (satu dari semua variabel prediktor yang tersedia) yang dapat memisahkan dataset pada simpul sebelumnya sehingga lebih homogen pada sub-simpul. Pemilihan variabel prediktor terbaik juga dilakukan dengan menggunakan indeks Gini.
3. Lakukan penyekatan dengan menggunakan variabel prediktor pada tahapan (2), periksa apakah sudah memenuhi kriteria pemberhentian, jika tidak maka lakukan kembali tahapan (1).

Random Forest

Random Forest merupakan pengembangan metode *Bagging (Bootstrap Aggregating)*. Metode ini bertujuan memperbaiki performa klasifikasi tunggal dengan nilai akurasi yang rendah (Wezel dan Potharst, 2007). Pengembangan metode *bagging* tersebut terletak pada proses resampling. Resampling pada metode *random forest* tidak hanya pada pengacakan observasinya, tetapi juga pada variabel prediktor, sehinggasetiap proses resampling yang dilakukan memuat sampel yang berbeda. Akibatnya ukuran dan bentuk pohon klasifikasi juga berbeda (Liaw dan Weiner, 2002). Berikut adalah tahapan *Random Forest* secara umum (Sartono dan Syafitri, 2010):

1. a. tahapan resampling

   Tahapan ini melakukan pembentukan sampel baru. Sampel dibentuk dari proses penarikan sampel acak dengan pemulihan sebanyak n data dari data training.

   b. tahapan random sub-setting

   Tahapan ini membentuk model pohon klasifikasi (salah satunya CART) dengan data yang telah diperoleh pada tahapan 1.a, namun setiap proses pemisahan dilakukan pemilihan secara acak m < d variabel penjelas. d adalah banyaknya variabel prediktor yang digunakan.

   c. ulangi langkah 1.a dan 1.b sebanyak B kali sehingga diperoleh B pohon klasifikasi (CART)

2. Majority Vote

Melakukan pendugaan gabungan berdasarkan hasil prediksi mayoritas dari B prediksi yang terbentuk.

***Extreme Gradient Boosting* (Xgboost)**

Extreme Gradien Boosting (Xgboost) merupakan metode *ensemble* yang mirip dengan *gradient boosting*, namun lebih efisien karena terdapat parameter regulasi yang dapat mengontrol kompleksitas model (Chen dan Guestrin, 2016). Secara sederhana tahapan dalam pembentukan model *gradient boosting* adalah sebagai berikut ((Friedman, 2000):

a. Membentuk model awal berupa nilai konstan yang memenuhi ketentuan berikut:

$$F_0(x) = argmin \sum_{i=1}^{n} L(y_i, \gamma) \qquad (2)$$

b. Untuk tahapan m = 1 sampai dengan M, maka lakukan: 1). Hitung residual atau error dari model sebelumnya untuk setiap observasi. 2). Pembentukan model pohon dengan menggunakan data x sebagai prediktor dan error pada tahapan sebelumnya sebagai respon. 3). Menghitung prediksi model. 4). Ulangi langkah 2.

c. pembentukan model akhir

$$\hat{y}_i^{(M)} = F_0(x) + \sum_{m=1}^{M} \gamma_m F_m(x) \qquad (3)$$

Algoritma *Xgboost* dapat digunakan dalam regresi maupun klasifikasi, dan dikenal memiliki performa yang baik dibanding metode *gradient boosting*.

Evaluasi Kebaikan Model

Kebaikan evaluasi dapat dianalisis dengan *Confusion Matrix* (Han, Kamber, dan Pei, 2012). Misalkan terdapat dua kelas data yaitu kelas positif dan kelas negatif, maka prediksi dapat diklasifikasikan sebagai berikut:

Tabel 1 Confusion Matrix

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

Sumber: Han, Kamber, dan Pei, 2012

Kebaikan model klasifikasi dapat diukur dengan formula berikut:

1. $Accuracy = \frac{\text{TP+TN}}{(\text{TP+TN+FN+FP})}$
2. $Sensitivity\ or\ recall = \frac{\text{TP}}{\text{TP+FN}}$
3. $Specificity = \frac{\text{TN}}{\text{TN+FP}}$
4. $Precision = \frac{\text{TP}}{\text{TP+FP}}$
5. $F - Score = \frac{2(\text{recall})(\text{ precision})}{\text{recall+precision}}$.

Evaluasi kebaikan model klasifikasi dapat juga dilihat dari nilai AUC, yaitu nilai yang menggambarkan luas area di bawah kurva ROC dengan nilai antara 0 hingga 1. Sedangkan, Kurva ROC adalah kurva yang menggambarkan kebaikan model klasifikasi yang disajikan dalam dua dimensi. Kurva tersebut memuat nilai persentase *False Positive* (1- *Specificity*) dengan persentase *True Positive* (Fawcett 2006).

Metode Analisis

Data yang digunakan dalam penelitian ini ialah data hasil Survei Sosial Ekonomi Nasional (SUSENAS) Kor tahun 2020 dan SUSENAS Modul Konsumsi 2020. Objek pengamatan dalam penelitian ini adalah rumah tangga. Variabel respon pada penelitian ini adalah status kemiskinan rumah tangga yang dikategorikan menjadi rumah tangga miskin jika pengeluaran konsumsi per kapita per bulan berada di bawah garis kemiskinan dan jika pengeluaran konsumsi per kapita per bulan melebihi garis kemiskinan maka rumah tangga dikategorikan rumah tangga tidak miskin. Per Maret 2020, nilai garis kemiskinan Provinsi Bengkulu adalah Rp 527.031 per bulan per orang.

Variabel prediktor yang digunakan ialah: Jenis Kelamin Kepala Rumah Tangga (X1), Usia Kepala Rumah Tangga (X2), Banyaknya Anggota Rumah Tangga (X3), Ijazah Tertinggi Kepala Rumah Tangga (X4), Status Pekerjaan Kepala Rumah Tangga (X5), Status Kepemilikan Rumah (X6), Bahan Bangunan Atap (X7), Bahan Utama Dinding Rumah (X8), Bahan Utama Lantai Rumah (X9), Penggunaan Fasilitas BAB (X10), Sumber Air Minum (X11), Bahan Bakar Utama Memasak (X12), Luas lantai (X13), dan Wilayah (X14).

Langkah –langkah analisis data yang dilakukan dalam penelitian ini adalah sebagai berikut:

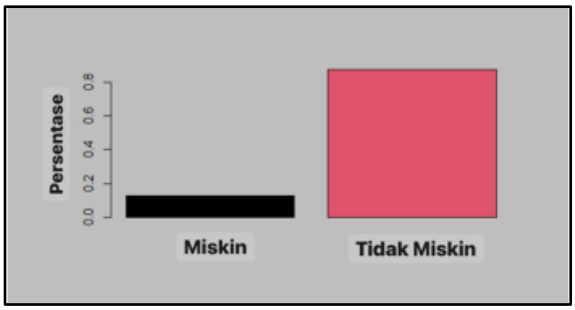
1. Pengklasifikasian status kemiskinan rumah tangga berdasarkan penghitungan garis kemiskinan per bulan per orang dengan jumlah anggota rumahtangga.
2. Eksplorasi data awal sehingga diperoleh gambaran umum dari data kemiskinan Provinsi Bengkulu, dengan menggunakan metode Khi-Kuadrat dan analisis Korelasi
3. Melakukan pengecekan data hilang (*missing data values*)dilakukan pengisian dengan data lain yang relevan, dan mengidentifikasi kondisi data (terkait proporsi antara rumah tangga miskin dan tidak miskin)
4. Membagi data menjadi data training dan testing: data training digunakan untuk pemodelan dan data testing digunakan untuk evaluasi performa klasifikasi. Data dibagi menjadi 75:25 dari total data yang tersedia.
5. Melakukan pemodelan klasifikasi dengan pohon klasifikasi tunggal (CART) dan pohon klasifikasi gabungan *Random Forest* serta *Xgboost* pada data training berdasarkan pembagian pada tahap 4.
6. Melakukan evaluasi performa klasifikasi menggunakan nilai-nilai pada confusion matrix hasil pemodelan pada tahap 5.
7. Melakukan perbaikan proporsi data pada kelas minoritas dengan menggunakan SMOTE, metode SMOTE diterapkan pada data training yang diperoleh dari tahap 4.
8. Melakukan pemodelan klasifikasi dengan

pohon klasifikasi tunggal (CART) dan pohon klasifikasi gabungan Random Forest serta Xgboost pada data training berdasarkan pembagian pada tahap 7.

9. Melakukan evaluasi performa klasifikasi menggunakan nilai-nilai pada confusion matrix hasil pemodelan tahap 8.
10. Memilih satu model terbaik berdasarkan nilai AUC, F-score, Sensitivity dan specificity
11. Mengidentifikasi variabel penting yang membangun model terbaik pada tahap 10.
12. Interpretasikan hubungan variabel penting dengan status rumah tangga.

HASIL DAN PEMBAHASAN

Pengklasifikasian rumah tangga yang tergolong miskin dan tidak miskin dilakukan berdasarkan garis kemiskinan Provinsi Bengkulu pada bulan Maret 2020. Total rumah tangga sampel adalah sebanyak 5730 rumah tangga, dengan rincian 721 rumah tangga terkategori miskin dan 5009 rumah tangga terkategori tidak miskin. Proporsi antara rumah tangga yang dikategorikan miskin jauh lebih kecil dibandingkan proporsi rumah tangga tidak miskin. Ketimpangan ini sering disebut sebagai kondisi data yang tidak seimbang (*unbalance dataset)*.



Gambar 2. Sebaran Status Kemiskinan Rumah Tangga di Provinsi Bengkulu.

Sumber: Data SUSENAS 2020 yang diolah.
Sumber: Data SUSENAS 2020 yang diolah

Terdapat 14 variabel penjelas yang digunakan pada penelitian ini yaitu Jenis Kelamin Kepala Rumah Tangga (X1), Usia Kepala Rumah Tangga (X2), Banyaknya Anggota Rumah Tangga (X3), Ijazah Tertinggi Kepala Rumah Tangga (X4), Status Pekerjaan Kepala Rumah Tangga (X5), Status Kepemilikan Rumah (X6), Bahan Bangunan Atap (X7), Bahan Utama Dinding Rumah (X8), Bahan Utama Lantai Rumah (X9), Penggunaan Fasilitas BAB (X10), Sumber Air Minum (X11), Bahan Bakar Utama Memasak (X12), Luas lantai (X13), dan Wilayah (X14). Untuk menentukan variabel prediktor yang memiliki hubungan dengan variabel respon yang digunakan yaitu jika variabel prediktor merupakan variabel numerik maka digunakan uji t-student, sedangkan untuk variabel kategorik digunakan uji Khi-kuadrat.

Tabel 2. Hasil pengujian Khi-Kuadrat/Fisher

Variabel	Uji khi-kuadrat
Jenis Kelamin	Signifikan
Pendidikan	Signifikan
kepemilikan rumah	Signifikan
jenis atap	Signifikan
jenis lantai	Signifikan
sumber air minum	Signifikan
bahan bakar	Signifikan
fasilitas BAB	Signifikan
pekerjaan KRT	Tidak Signifikan
Wilayah	Signifikan

Sumber: Data SUSENAS 2020 yang diolah

Tabel 3. Hasil pengujian t-student

Variabel	Uji-t
Luas lantai	Signifikan
Usia	Signifikan
Jumlah anggota rumah tangga	Signifikan

Sumber: Data SUSENAS 2020 yang diolah

Pemodelan Klasifikasi Kemiskinan

a. Model menggunakan data awal

Analisis klasifikasi kemiskinan dilakukan menggunakan model klasifikasi pohon tunggal dan gabungan. Sebelum dilakukan analisis, dataset dibagi menjadi data *training* dan data *testing.* Data *training* digunakan untuk membangun model pembelajaran, kemudian hasil model akan dievaluasi menggunakan data *testing*. Pembagian dataset dilakukan secara acak dengan proporsi 75%:25%. Berikut adalah gambaran pembagian data *training* dan *testing*:

Tabel 4. Pembagian Data

	Total	Miskin	Tidak Miskin
Data Training	4296	540	3756
Data Testing	1434	181	1253

Sumber: Data SUSENAS 2020 yang diolah

Pemodelan menggunakan metode klasifikasi CART dilakukan dengan terlebih dahulu mencari parameter optimal yaitu *cost complexity*, kedalaman pohon dan minimal observasi dalam setiap simpul pohon. Berikut rincian beberapa kombinasi parameter yang menghasilkan performa klasifikasi paling baik:

Tabel 5. Kombinasi Parameter Optimal model CART

No	Cost	Keda-laman	Minsplit	AUC
1	**$2.5x10^{-5}$**	**13**	**22**	**0.751**
2	$8{,}2x10^{-4}$	14	14	0.747
3	$1.87x10^{-6}$	7	13	0.747

Sumber: Data SUSENAS 2020 yang diolah

Berdasarkan Tabel 5. dapat dilihat bahwa terdapat kombinasi parameter yang memberikan nilai AUC paling tinggi. Selanjutnya akan dilakukan pemodelan klasifikasi CART dengan menggunakan cost sebesar 2.5x10-5, kedalaman pohon sebesar 13 dan minimal observasi dalam simpul sebesar 22. Jika ditinjau dari kedalaman pohon, model yang dihasilkan cukup kompleks.

Sama halnya dengan pemodelan CART, pada *Random Forest* juga terdapat beberapa parameter yang harus ditetapkan diawal pemodelan. parameter tersebut antara lain adalah mtry (jumlah variabel prediktor yang diambil secara acak dari total variabel yang tersedia), ntree (banyak pohon yang dibentuk) dan minsplit (minimal observasi pada setiap simpul). Berikut adalah beberapa kombinasi dari parameter dalam pemodelan random forest yang memberikan performa paling baik:

Tabel 6. Kombinasi Parameter Optimal model *Random Forest*

No	Mtry	Ntree	Minsplit	AUC
1	**12**	**1270**	**40**	**0.796**
2	12	691	36	0.794
3	12	1657	31	0.793

Sumber: Data SUSENAS 2020 yang diolah

Berdasarkan Tabel 6 diperoleh model *random forest* terbaik dihasilkan jika pemodelan dilakukan menggunakan parameter mtry sebesar 12, ntree sebesar 1270 dan minsplit sebesar 40. Ntree sebesar 1270 mengindikasikan terdapat 1270 pohon klasifikasi tunggal yang terbentuk pada pemodelan.

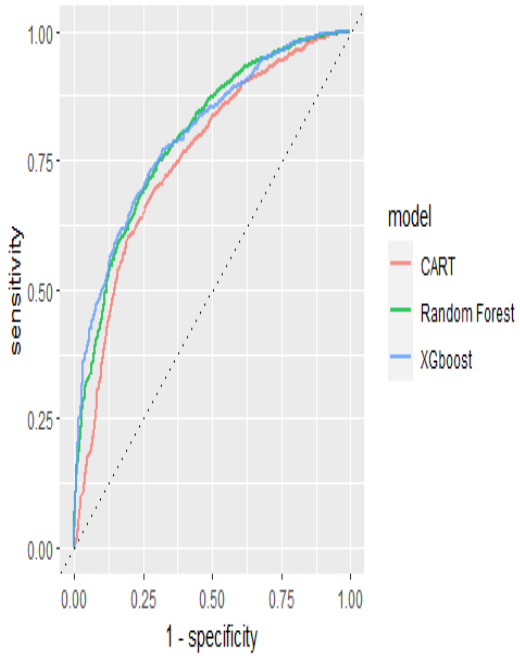
Pemodelan klasifikasi menggunakan *Xgboost* memiliki parameter yang mirip dengan *Random Forest* yaitu mtry, ntree dan minsplit. Tabel 6 melampirkan kombinasi terbaik dari parameter ini, dengan nilai AUC paling tinggi jika pemodelan *Xgboost* dilakukan menggunakan mtry sebesar 3, ntree sebesar 413 dan minsplit sebesar 24.

Tabel 7. Kombinasi Parameter Optimal model XGboost

No	Mtry	Ntree	Minsplit	AUC
1	**3**	**413**	**24**	**0.800**
2	3	864	19	0.791
3	3	1359	35	0.790

Sumber: Data SUSENAS 2020 yang diolah

Gambar 3 menyajikan perbandingan performa klasifikasi CART, *Random Forest* dan *Xgboost* pada data yang tidak seimbang (proporsi rumah tangga miskin jauh lebih sedikit). Dapat dilihat bahwa pemodelan dengan *Xgboost* berada paling atas, meskipun memiliki nilai yang tidak terlalu berbeda dengan *Random Forest*. Dari ketiga model, pemodelan dengan pohon klasifikasi tunggal (CART) memiliki performa klasifikasi yang paling rendah.



Gambar 3. Kurva ROC CART, *Random Forest* dan *Xgboost*

Sumber: Data SUSENAS 2020 yang diolah

Tabel 8. Evaluasi Model Klasifikasi

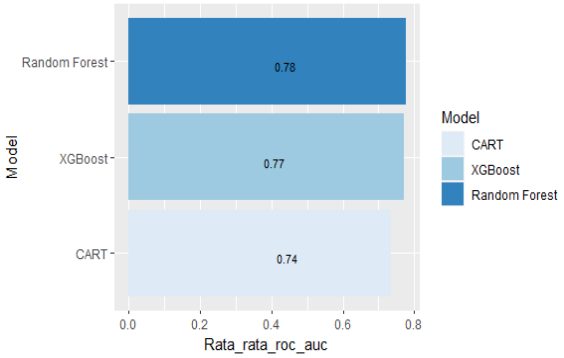
Model	F-score	A	B	AUC
CART	0.177	0.122	0.964	0.749
Random Forest	0.148	0.088	0.985	0.804
XGboost	**0.261**	**0.182**	**0.969**	**0.801**

Sumber: Data SUSENAS 2020 yang diolah

dimana A ialah Sensitivity dan B ialah *Specivicity*. Tabel 8. Merupakan hasil evaluasi setiap pemodelan klasifikasi yang digunakan dalam penelitian ini. performa klasifikasi dinilai dari hasil prediksi pada data testing. Dikarenakan data memiliki permasalahan ketidakseimbangan data, maka performa klasifikasi dapat dibandingkan dengan melihat nilai F-score. Nilai F-score tertinggi diperoleh pada pemodelan *Xgboost*, namun pemodelan *Xgboost* tidak cukup baik dalam mengatasi permasalahan ketidakseimbangan data. Hal ini dapat dilihat dari nilai sensitivity sebesar 0.182, yang berarti metode ini hanya mampu mengklasifikasikan rumah tangga miskin secara tepat sebasar 18.2%. Sehingga perlu penanganan permasalahan ini pada level data yaitu dengan metode SMOTE.

b. Model menggunakan data setelah SMOTE

Pemodelan CART pada data setelah SMOTE memiliki performa paling baik saat *cost_complexity* sebesar 2,74 x 10-7, kedalaman pohon sebesar 6 dan minsplit 25. Pemodelan Random Forest memliki performa paling baik saat ntree sebanyak 1270, mtry sebesar 12 dan minsplit sebesar 25. Sedangkan pada Xgboost, performa paling baik saat ntree sebanyak 413, mtry sebesar 3 dan minsplit sebesar 24. Gambar 4 menyajikan perbandingan rata-rata nilai AUC yang diperoleh pada setiap model.



Gambar 4. Rata-rata AUC pada Setiap Model

Sumber: Data SUSENAS 2020 yang diolah

Selanjutnya, setiap pemodelan yang dihasilkan dievaluasi performa klasifikasinya dengan melihat nilai F-score, *sensitivity*, *specificity* dan AUC yang disajikan dalam Tabel 9. Evaluasi dilakukan dengan menerapkan setiap model pada data testing.

Tabel 9. Evaluasi Model Klasifikasi

Model	F-score	A	B	AUC
CART	0.344	0.569	0.749	0.736
Random Forest	**0.365**	**0.409**	**0.879**	**0.791**
XGboost	0.350	0.376	0.888	0.779

Sumber: Data SUSENAS 2020 yang diolah

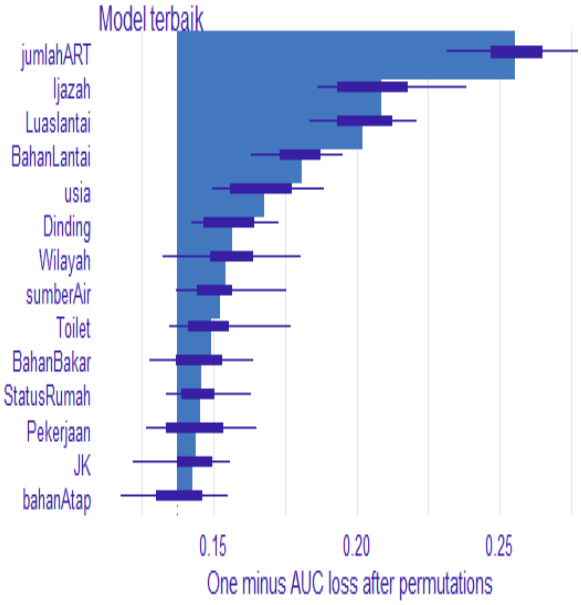
dimana A ialah Sensitivity dan B ialah *Specivicity*

Berdasarkan Tabel 9 dapat dilihat bahwa model Random Forest memiliki nilai AUC paling tinggi dan nilai sentivity tidak terlalu rendah yaitu 0.409 dan specivicity 0.879 Sehingga dalam interpretasi model akan digunakan Random Forest sebagai

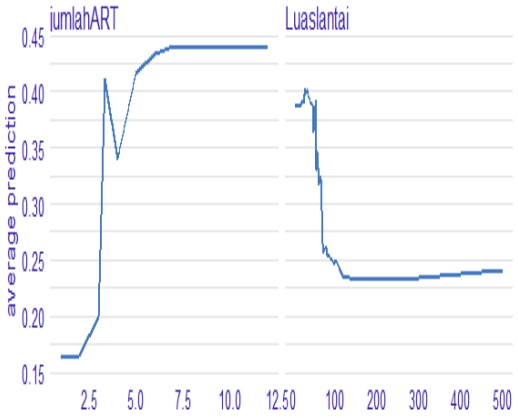
identifikasi karakteristik rumah tangga miskin dan tidak miskin.

Variable Importance

Variable Importance digunakan untuk melihat tingkat kepentingan setiap variabel yang digunakan di dalam pemodelan yang terpilih pada tahapan sebelumnya. Dari 14 variabel prediktor, terdapat tiga variabel prediktor yang memiliki kontribusi paling tinggi dalam mengklasifikasikan rumah tangga miskin dan tidak miskin, yaitu : jumlah anggota rumah tangga, ijazah terakhir yang dimiliki kepala keluarga dan luas lantai rumah tempat tinggal (Gambar 5). Hubungan antara variabel tersebut terhadap rumah tangga miskin dapat dilihat dari Gambar 6. Hubungan antara jumlah anggota rumah tangga dengan rumah tangga miskin adalah positif, yang mengindikasikan bahwa semakin banyak anggota rumah tangga akan semakin besar peluang rumah tangga tersebut masuk kedalam kategori miskin. Sedangkan luas lantai rumah memiliki hubungan negatif yang berarti semakin sempit rumah yang ditempati maka semakin besar pelua rumah tangga tersebut terkategori miskin.



Gambar 5. *Variable Importance* Model Random Forest pada data SMOTE
Sumber: Data SUSENAS 2020 yang diolah



Gambar 6. Partial Dependence Plot Variabel Jumlah ART dan Luas Lantai.

Sumber: Data yang diolah

KESIMPULAN DAN SARAN

Berdasarkan hasil dan pembahasan yang diuraikan pada bab sebelumnya, dapat diambil kesimpulan bahwa: Pemodelan dengan pohon klasifikasi pohon gabungan dalam hal ini Random Forest dan Xgboost tidak dapat memberikan performa yang baik pada data yang tidak seimbang. Penerapan metode SMOTE pada data dapat meningkatkan ketepatan klasifikasi pada kelas minoritas (rumah tangga miskin), namun belum maksimal. Terdapat tiga variabel yang paling berkontribusi besar dalam pengklasifikasian dengan metode SMOTE+Random Forest yaitu jumlah anggota rumah tangga, ijazah terakhir kepala rumah tangga dan luas lantai rumah. Temuan ini diharapkan dapat digunakan sebagai dasar analisis pengambilan keputusan, khususnya untuk mengidentifikasi rumahtangga miskin, sehingga program-program bantuan untuk penganggulangan kemiskinan diharapkan lebih tepat sasaran di masa mendatang.

Pada penelitian selanjutnya, metode penanggulangan ketidakseimbangan data lainnya sebaiknya juga dilakukan, sehingga ketepatan klasifikasi lebih tinggi pada kelas minoritas.

DAFTAR PUSTAKA

Badan Pusat Statistik (BPS). 2020. Diakses pada 29 Maret 2021 melalui: https://www.bps.go.id/pressrelease/2021/02/15

/1851/persentase-penduduk-miskin-september-2020-naik-menjadi-10-19-persen.html
Breiman L., Friedman J.H., Olshen R.A., Stone C.J.1984. Classification and Regression Trees. New York: Chapman & Hall/CRC.
Chawla N.V., Browyer K.W., Hall L.O., Kegelmeyer W.P. 2002. SMOTE : Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 16, 321-357.
Chen, T., & Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
Han, J., Pei, J., & Kamber, M. 2012. Data mining: concepts and techniques. Elsevier.
Fawcett T. 2006. An introduction to ROC analysis. Pattern Recognition Letters. 27, 861–874.
Ishartono dan Raharjo S.T. 2016. Sustainable Development Goals (SDGs) dan Pengetasan Kemiskinan. Share: Social Work Jurnal, Vol 6, 2
Ispriyanti, D., Prahutama, A., dan Mustafid. 2019. Analisis Klasifikasi Kemiskinan Kota Semarang Menggunakan Algoritma QUEST. Statistika, Vol.7(1)
Lewis, R. J. 2000. An introduction to classification and regression tree (CART) analysis. In Annual meeting of the society for academic emergency medicine in San Francisco, California (Vol. 14).
Liaw A.,Wiener M. 2002. Classification and Regression. R News, 2,18-22
Sartono, B. dan Syafitri., T. U. 2010. Metode Pohon Gabungan: Solusi Pilihan Untuk Mengatasi Kelemahan Pohon Regresi Dan Klasifikasi Tunggal. Forum Statistika dan Komputasi, 1-7
Sutton C. D. 2005. Classification and Regression Trees, Bagging, and Boosting. Handbook of Statistics, 24, 303-329.
Wezel M.V., Potharst R. 2007. Improved Customer Choice Predictions using Ensemble Methods. European Journal of Operational Research, 181, 43