Jurnal Teknik Informatika (JUTIF)
Vol. 2, No. 1, Juni 2021, hlm. 51-56

DOI: https://doi.org/10.20884/1.jutif.2021.2.1.52
p-ISSN: 2723-3863
e-ISSN: 2723-3871

COMPARISON OF MACHINE LEARNING METHODS IN CLASSIFYING POVERTY IN INDONESIA IN 2018

Pardomuan Robinson Sihombing*[1], Ade Marsinta Arsani[2]

[1,2] Badan Pusat Statistik, Jakarta Pusat, Indonesia
Email: [1]robinson@bps.go.id, [2]ade.marsinta@bps.go.id

(Naskah masuk: 11 Januari 2021, diterima untuk diterbitkan: 15 Januari 2021)

Abstract

Poverty is still one of the main problems in economic development besides inequality, unemployment, and economic growth. This study aims to model poverty directly using a discrete choice model, namely the machine learning classification method. The data used are imbalanced data where one of the categories is small enough so that the resample of both sampling method is used. In this study, several machine learning methods were applied, including the Decision Tree, Naïve Bayes, K-Nearest Neighbor (KNN), and Rotation Forest. The results show that the technique of using resample both samplings provides optimal results for the four machine learning methods. If viewed from the indicators of accuracy, specificity, sensitivity, AUC, and the highest Kappa coefficient produced, the best method is the KNN method. The KNN model has an accuracy value of 0.73 percent, sensitivity of 0.68 percent, specificity of 78 percent, and AUC of 0.73.

Keywords: *decision tree, K-Nearest Neighbour, machine learning, naïve bayes, rotation forest.*

PERBANDINGAN METODE *MACHINE LEARNING* DALAM KLASIFIKASI KEMISKINAN DI INDONESIA TAHUN 2018

Abstrak

Kemiskinan masih menjadi salah satu masalah pokok dalam pembangunan ekonomi selain ketimpangan, pengangguran dan pertumbuhan ekonomi. Penelitian ini bertujuan memodelkan kemiskinan secara langsung dengan menggunakan model pilihan diskrit yaitu metode pengklasifikasian model *machine learning*. Data yang digunakan merupakan data yang *imbalanced* dimana salah satu kategori nilainya cukup kecil sehingga digunakan metode *resample both sampling*. Dalam penelitian ini menerapkan beberapa metode *machine learning* di antaranya *Decision Tree*, *Naïve Bayes*, *K-Nearest Neighbour* (KNN) dan *Rotation Forest*. Hasil yang didapat bahwa teknik menggunakan *treatment resample both sampling* memberikan hasil yang optimal untuk keempat metode *machine learning*. Jika dilihat dari indikator *accuracy*, *specificity*, *sensitivity*, AUC dan koefisien Kappa tertinggi yang dihasilkan, maka metode terbaik adalah metode KNN. Adapun model KNN memiliki nilai a*ccuracy* 0.73 persen, *sensitivity* sebesar 0.68 persen, s*pecificity* sebesar 78 persen, dan AUC 0.73.

Kata kunci: *decision tree, K-Nearest Neighbour, machine learning, naïve bayes, rotation forest*.

1. PENDAHULUAN

Kemiskinan masih menjadi salah satu masalah pokok dalam pembangunan ekonomi selain ketimpangan, pengangguran dan pertumbuhan ekonomi. Salah satu pilar *Sustainable Development Goals* (SDGs) dalam bidang pembangunan sosial adalah tercapainya pemenuhan hak dasar manusia yang berkualitas secara adil dan setara untuk meningkatkan kesejahteraan bagi seluruh masyarakat, dengan prioritas utama mengakhiri kemiskinan dalam segala bentuk di manapun (*end poverty in all its forms everywhere*). Data persentase kemiskinan Indonesia sendiri menunjukkan *trend* yang menurun dari tahun ke tahun, dimana angka kemiskinan tahun 2019 berdasarkan data Susenas Maret 2019 sebesar 9,41 persen.

Pada dasarnya ada dua pendekatan dalam memodelkan faktor yang mempengaruhi kemiskinan. Pendekatan pertama dengan menggunakan pendekatan regresi antara pengeluaran konsumsi per *adult equivalent* terhadap sejumlah variabel penjelas yang potensial yang disebut sebagai pendekatan konsumsi. Pendekatan kedua yang dapat dilakukan adalah dengan memodelkan kemiskinan secara langsung dengan menggunakan model pilihan diskrit. Pendekatan diskrit yang dimaksud adalah mengkategorikan kemiskinan menjadi dua kategori berdasarkan pengeluaran konsumsi rumah tangga dibandingkan garis kemiskinan suatu wilayah [1].

Menurut Han dan Kamber [2], klasifikasi adalah proses untuk menemukan model atau fungsi yang dapat menggambarkan dan membedakan kelas data atau konsep, dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas yang belum diketahui dari suatu objek pengamatan.

Beberapa metode klasifikasi kasik yang digunakan adalah analisis diskriminan dan regresi logistik. Dengan semakin pesatnya data dalam jumlah besar yang dikenal dengan data besar (*big data*), maka dibutuhkan metode-metode analisis yang dapat mengolah data besar tersebut dengan tepat dan cepat. Di sisi lain, di tengah pesatnya perkembangan teknologi kecerdasan buatan atau *artificial intelligence* (AI) maka berkembangnya metode yang dikenal dengan *machine learning*. Teknologi *machine learning* (ML) merupakan mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya. Pembelajaran mesin dikembangkan berdasarkan disiplin ilmu lainnya seperti statistika, matematika dan *data mining* sehingga mesin dapat belajar dengan menganalisa data tanpa perlu di program ulang atau diperintah. Metode *machine learning* yang dapat digunakan dalam pengklasifikasian diantaranya *Decition Trees* (DT), *Naive Bayes* (NB), *Random Forest*, *Rotation Forest* (RF), *K-Nearest Neighbor* (KNN), *Artificial Neural Networks* (ANN), *Support Vector Machine* (SVM), dan lainnya.

Masing-masing metode memiliki kelebihan dan kekurangan masing-masing. Metode *Desicion Tree* mampu mengintegrasikan model yang mudah ke dalam sistem basis data serta memiliki akurasi yang baik serta dapat menemukan kombinasi data yang tidak terduga. Dalam penelitian Kaunang [3] *decition tree* memberikan performa hingga 80 persen dalam mengklasifikasikan kemiskinan. Sementara itu *Rotation Forest* mampu memperbaiki kemampuan prediksi pada decision tree [4]. *Naïve Bayes classification* terbukti memiliki akurasai dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* dengan data yang besar [5]. Annur [6] menggunakan *naïve bayes* dalam mengklasifikasikan kemiskinan di Gorantalo dan mendapat hasil akurasi hingga 73 persen. Algoritma KNN memiliki beberapa kelebihan yaitu ketangguhan terhadap training data yang memiliki banyak *noise* dan efektif apabila *training* datanya besar [7]. Kurnia [8] mengklasifikasikan kemiskinan menggunakan metode KNN dengan tingkat akurasi hingga 90 persen.

Pada umumnya dalam metode pengklasifikasian ini didasarkan pada asumsi bahwa banyaknya data terdistribusi secara merata antara kelas yang berbeda. Padahal dalam kehidupan nyata, terdapat peristiwa yang menunjukkan bahwa banyaknya data yang tidak seimbang antara kelas yang berbeda, dan ketika kondisi ketidakseimbangan masalah ini disebut *imbalanced data* [9]. Menerapkan fungsi tautan yang tidak fleksibel ke data dengan fitur khusus ini mungkin mengakibatkan kesalahan spesifikasi tautan. Konsekuensi kesalahan spesifikasi tautan telah dipelajari oleh sejumlah penulis dalam literatur. Khususnya, untuk pengamatan biner independen, Czado dan Santner [10] menunjukkan bahwa dengan salah mengasumsikan tautan logistik mengarah ke peningkatan substansial dalam bias dan *mean squared error* dari estimasi parameter serta probabilitas yang diprediksi, baik secara asimptotik maupun dalam sampel terbatas. King dan Zeng [11] menyatakan bahwa ketika metode klasifikasi digunakan pada kasus *imbalanced* data, maka pengklasifikasian cenderung menihilkan peluang dari kelas minoritas karena nilai prediksi akan cenderung pada kelas mayoritas, sehingga tingkat ketepatan klasifikasi yang dihasilkan menjadi kurang baik. Terutama data yang digunakan adalah data dalam jumlah yang besar (*big data*). Purwa [12] melakukan perbandingan beberapa metode *resample* dan di dapat hasil bahwa metode *both/combine sampling* menghasilkan performa yang terbaik.

Oleh karena itu, pada penelitian ini mengkaji dan menerapkan beberapa metode *machine learning* seperti DT, NB, KNN dan RF dengan memperhatikan *imbalanced* data dan set data besar. Skema yang digunakan adalah menggunakan pembagaian data dengan metode deterministik (*holdout*) dengan melakukan *resample* kombinasi *undersampling* dan *oversampling* sekaligus (both/ *combine sampling*) dalam pemodelan klasifikasi status miskin rumah tangga di Indonesia.

2. METODE PENELITIAN

2.1. Sumber Data dan Variabel Penelitian

Sumber data yang digunakan dalam penelitian ini berasal dari data Susenas Maret 2018. Jumlah sampel yang digunakan sebanyak 259.378 rumah tangga. Adapun variabel penelitian yang digunakan dapat ditunjukkan oleh tabel 1.

Tabel 1. Variabel Penelitian

Variabel	Nama	Keterangan
Y	Status Miskin	0 Tidak Miskin 1 Miskin
X1	Tipe Daerah	0 Perkotaan 1 Perdesaan
X2	Status Perkawinan	0 Belum Kawin dan Cerai 1 Kawin
X3	Jenis Kelamin	0 Pria 1 Wanita
X4	Pendidikan	0 Tidak Sekolah 1 SD dan SMP 2 SMA 3 Perguruan Tinggi (PT)
X5	Lapangan Usaha	0 Primer 1 Sekunder 2 Tersier
X6	Jumlah ART	
X7	Umur	

2.2. *Modelling*

Pada tahap *modeling* akan dilakukan pembentukan model yang dapat membedakan kelas data. Untuk melakukan *modeling*, pada tahap ini dibutuhkan dua jenis data set yang berasal dari data hasil preprocessing yaitu *training* data dan *testing* data. *Training* data merupakan dataset yang digunakan untuk membangun model sementara *testing* data digunakan untuk menghitung *performance* dari model yang terbentuk dengan membandingkan label data sebenarnya dan label data hasil klasifikasi model. Untuk membentuk *training* data dan t*esting* data penelitian menggunakan metode deterministik/ *holdout*. Pembagian data set menjadi *training* data dan *testing* data dapat menggunakan cara deterministik/holdout, yaitu dengan menentukan sendiri rasio pembagian dari kedua dataset tersebut. Contohnya rasio dataset dapat menggunakan 8:2, artinya 0.8 dari keseluruhan data digunakan untuk *training* data dan 0.2 sisanya digunakan untuk *testing* data yang dapat menghasilkan *performances*.

2.3. Decision Tree

Decision Tree adalah struktur *flowchart* yang menyerupai *tree* (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun merepresentasikan kelas atau distribusi kelas. Alur pada *decision tree* di telusuri dari simpul akar ke simpul daun yang memegang prediksi. Dalam membangun sebuah *decision tree* secara *topdown* (dari atas ke bawah), tahap awal yang dilakukan adalah mengevaluasi semua atribut yang ada menggunakan suatu ukuran statistik (yang biasa digunakan adalah *information gain*) untuk mengukur efektifitas suatu atribut dalam mengklasifikasikan suatu kumpulan sampel data. Atribut yang diletakkan pada *root_node* adalah atribut yang memiliki *information gain* terbesar. Semua atribut adalah bersifat kategori yang bernilai diskrit. Atribut dengan nilai *continuous* harus didiskritkan [13].

2.4. *Naïve Bayes*

Naïve bayes classification adalah pengklasifikasian statistik yang dapat digunakan untuk memprediski probabilitas keanggotaan suatu kelas. *Naïve bayes classification* didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi. Metode Bayes merupakan pendekatan statistik untuk melakukan inferensi induksi pada persoalan klasifikasi. Pertama kali dibahas terlebih dahulu tentang konsep dasar dan definisi pada Teorema Bayes, kemudian menggunkan teorema ini untuk melakukan klasifikasi dalam *data mining*. Teorema Bayes memiliki bentuk umum seperti persamaan 1.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Keterangan:
X = Data dengan kelas yang belum diketahui
H = Hipotesis data X merupakan suatu class spesifik
P(H|X) = Probabilitas hipotesis H berdasarkan kondisi x (*posteriori prob.*)
P(H) = Probabilitas hipotesis H (*prior prob.*)
P(X|H) = Probabilitas X berdasarkan kondisi tersebut
P(X) = Probabilitas dari X

2.5. K-*Nearest Neighbor*

Algoritma *K-Nearest Neighbor* (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran. Sebuah titik pada ruang ini ditandai kelas c jika kelas c merupakan klasifikasi yang paling banyak ditemui pada k buah tetangga terdekat titk tersebut. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak Euclidean dengan rumus seperti pada persaman 2 :

$$distance = \sqrt{\sum_{i=1}^{n}\left(X_{training}^{i} - X_{testing}\right)^2} \quad (2)$$

Pada fase pembelajaran, algoritma ini hanya melakukan penyimpanan vektor-vektor fitur dan klasifikasi dari data pembelajaran. Pada fase klasifikasi, fitur-fitur yang sama dihitung untuk data *test* (yang klasifikasinya tidak diketahui). Jarak dari vektor yang baru ini terhadap seluruh vektor data pembelajaran dihitung, dan sejumlah K buah yang paling dekat diambil. Titik yang baru klasifikasinya diprediksikan termasuk pada klasifikasi terbanyak dari titik-titik tersebut. Nilai K yang terbaik untuk algoritma ini tergantung pada data. Secara umum, nilai K yang tinggi akan mengurangi efek *noise* pada klasifikasi, tetapi membuat batasan antara setiap klasifikasi menjadi lebih kabur. Nilai K yang bagus dapat dipilih dengan optimasi parameter, misalnya dengan menggunakan *cross-validation*. Kasus khusus di mana klasifikasi diprediksikan berdasarkan data pembelajaran yang paling dekat (dengan kata lain, K = 1) disebut algoritma *nearest neighbour*.

2.6. *Rotation Forest*

Rotation forest adalah metode pohon gabungan yang menggunakan prinsip komponen utama untuk menyusun pohon keputusannya. Analisis komponen utama digunakan untuk merotasi sumbu peubah yang akan dibangun pohon keputusannya. Meskipun menggunakan analisis komponen utama, semua komponen utama tetap digunakan untuk membangun pohon keputusan agar menjaga keragaman informasi data [14]. Misalkan $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ adalah poin data

sebanyak n peubah dan $\boldsymbol{X}$ adalah gugus data yang terdiri dari data latih dalam bentuk matriks. Misalkan $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ adalah sebuah vektor dengan label kelas pada data. Pengklasifikasi dalam metode ini dilambangkan dengan $D_1, \ldots, D_L$ dan $\boldsymbol{F}$ sebagai gugus peubah. Seperti metode klasifikasi lainnya, dalam *rotation forest* perlu ditentukan jumlah pohon yang akan dibangun yaitu sebanyak L lalu seluruh pengklasifikasi dapat dilatih secara bersama.

2.8. Evaluasi

Evaluasi dilakukan untuk memilih metode pembagian data set dan klasifikasi yang mana yang dapat menghasilkan tingkat akurasi. Evaluasi dalam penelitian adalah dengan memperhatikan *Confussion Matrix*. *Confussion Matrix* merupakan sebuah alat untuk mengetahui sejauh mana pengklasifikasian dapat mengenal atau memprediksi kelas data. *Confussion Matrix* merupakan tabel berukuran *mxm* dengan *m*=jumlah kelas [15]. Bagian kolom diisi oleh label aktual untuk tiap kelas, sementara bagian baris diisi oleh label kelas hasil prediksi seperti tabel 2.

Tabel 2. Confusion Matrix

Confusion Matrix		*Actual Class*		*Total*
		Yes	*No*	
Predicted Class	*Yes*	*TP*	*FP*	*P'*
	No	*FN*	*TN*	*N'*
Total		*P*	*N*	

Pada umumnya ketepatan pengklasifikasian digunakan ukuran akurasi yaitu proporsi frekunsi yang tepat diklasifikasikan dengan total sampel yang ada. Selain melihat akurasi kita dapat melihat sensitivity. *Sensitivity* + merupakan proporsi kelas yang menjadi perhatian/diinginkan terprediksi dengan benar. *Specificity* - merupakan proporsi kelas yang tidak menjadi perhatian/tidak diinginkan terprediksi dengan benar. Apabila tingkat akurasi tinggi, namun *sensitivity* dan *specificity* rendah, maka pengklasifikasian dapat dikatakan tidak baik.

$$accuracy = \frac{TP+TN}{P+N} \quad (3)$$

$$Sensitivity+= \frac{TP}{P} \quad (4)$$

$$Specificity-= \frac{TN}{N} \quad (5)$$

Ukuran evaluasi kinerja klasifikasi lain adalah kurva *Receiver Operating Characteristic* (ROC). Kurva ROC adalah kurva analisis yang menggambarkan kinerja suatu model klasifikasi pada dua dimensi antara *sensitivity* sebagai sumbu y dan (1-*spesificity*) sebagai sumbu x [16]. Nilai tunggal yang dapat digunakan untuk mengukur kinerja klasifikasi pada kurva ROC adalah *Area Under Curve the ROC* (AUC). Tabel 3 menunjukkan standar kategori pengklasifikasian berdasarkan nilai AUC [17].

Selain itu, kebaikan model dapat dilihat dengan nilai Kappa dimana nilai yang dipakai untuk menentukan kekuatan kesepakatan/reliabilitas. Tes diagnostik ini juga yang dianjurkan oleh Landis dan Koch [18] semakin tinggi nilai Kappa akan semakin baik model yang digunakan.

Tabel .3 Kategori klasifikasi berdasarkan nilai AUC

Nilai AUC	Kategori Pengklasifikasian
0.90 – 1.00	*Excellent classification*
0.80 - 0.90	*Good classification*
0.70 – 0.80	*Fair classification*
0.60 – 0.70	*Poor classification*
0.50 – 0.60	*Failure*

2.9. Tahapan dalam Analisis Data

Adapun tahapan dalam melakukan analisis data adalah sebagai berikut:

- Melakukan *resample* terhadap data yang ada dengan menggunakan teknik *combine/ both sampling*
- Melakukan *modeling* data dengan teknik deterministik yaitu 80 persen untuk data *training* dan 20 persen untuk data *testing*
- Me*running* keempat metode *machine learning* (*Decision Tree*, *Naïve Bayes*, *K-Nearest Neighbour* dan *Rotation Forest*)
- Melakukan evaluasi terhadap keempat metode
- Memilih metode terbaik berdasarkan kriteria *accuracy*, *specificity*, *sensitivity*, AUC dan koefisien Kappa tertinggi
- Dalam pengolahan data menggunakan *software* R versi 4.0

3. HASIL DAN PEMBAHASAN

Sebelum melakukan analisis lebih lanjut mengenai hubungan antar variabel, dilakukan analisis deskriptif mengenai variabel penelitian. Dari sampel yang ada terdapat 10.79 persen penduduk Indonesia yang masih di bawah garis kemiskinan. Perbedaan fasilitas dan infrastuktur di daerah perkotaan dan pedesaan dapat mempengaruhi tingkat kesejaterahan suatu rumah tangga. Sebanyak 8.50 persen penduduk miskin Indonesia berada di perdesaan. Jika dilihat dari status perkawinannya, didominasi penduduk yang memiliki pasangan sebesar 9.49 persen. Sedangkan jika dilihat dari jenis kelamin kepada rumah tangga (KRT), penduduk miskin didominasi oleh KRT yang dikepalai seorang pria. Analisis deskriptif variable penelitian dapat ditunjukkan oleh tabel 4.

3.1. Pemilihan Model

Hasil klasifikasi model pada data *imbalanced* tanpa dilakukan *treatment* apapun dengan menggunakan keempat model *machine learning* secara umum menghasilkan nilai *accuracy* dan *sensitivity* yang lebih tinggi jika dibandingkan dengan data yang sudah di-*treatment* dengan menggunakan *resample combine / both sampling* seperti yang terlihat pada Tabel 5. Akan tetapi di sisi lain, performa klasifikasi tanpa *treatment* menghasilkan

specificity, AUC dan Kappa yang lebih rendah jika dibandingkan dengan data *resample* pada keempat model seperti yang terlihat pada Tabel 6.

Tabel 4 Analisis Deskriptif Variabel Penelitian

Variable	Poverty Status	Tidak Miskin	Miskin	Total
Tipe Daerah	Perdesaaan	50.63	8.5	59.13
	Perkotaan	38.58	2.29	40.87
Status Perkawinan	Belum Kawin	13.37	1.3	14.67
	Kawin	75.84	9.49	85.33
Jenis Kelamin	Pria	79.24	9.61	88.85
	Wanita	9.97	1.18	11.15
Sektor Lapangan Usaha	Primer	37.7	7.45	45.16
	Sekunder	15	1.39	16.39
	Tersier	36.5	1.95	38.45
Pendidikan Tertinggi	Tidak Sekolah	3.95	1.17	5.12
	SD dan SMP	52.37	7.97	60.34
	SMA	23.64	1.48	25.12
	PT	9.25	0.18	9.42
Total		89.21	10.79	100

Tabel 5. Rata-rata Performa Klasifikasi Machine Learning Berdasarkan *Accuracy* dan *Sensitivity*

Model	*Treatment*	*Accuracy*	*Sensitivity*
Decision Tree	*No Treatment*	0.89	0.89
	Both Sampling	0.69	0.69
Naïve Bayes	*No Treatment*	0.89	0.9
	Both Sampling	0.68	0.7
KNN	*No Treatment*	0.89	0.99
	Both Sampling	0.73	0.68
Rotation Forest	*No Treatment*	0.89	1.00
	Both Sampling	0.70	0.70

Tabel 6. Rata-rata Performa Klasifikasi Machine Learning Berdasarkan *Specificity*, AUC dan nilai Kappa

Model	*Treatment*	*Specificity*	AUC	Kappa
Decision Tree	*No Treatment*	0.00	0.50	0.00
	Both Sampling	0.70	0.73	0.38
Naïve Bayes	*No Treatment*	0.41	0.56	0.17
	Both Sampling	0.67	0.68	0.37
KNN	*No Treatment*	0.06	0.52	0.07
	Both Sampling	0.78	0.73	0.46
Rotation Forest	*No Treatment*	0.00	0.50	0.00
	Both Sampling	0.30	0.76	0.40

Jika dilihat dari nilai AUC, dengan skema tanpa *treatment* dengan keempat model menghasilkan nilai yang relatif sama, yaitu berkisar antara 50 persen sampai 56 persen. Nilai rata-rata *accuracy* dan *sensitivity* tertinggi yang mampu dihasilkan oleh data *imbalanced* tanpa *treatment* masing-masing sebesar 89 persen dan 100 persen dengan teknik *Rotation Forest*. Sedangkan rata-rata nilai *specificity* yang dihasilkan hanya sebesar 11 persen. Jika dilihat nilai yang cukup kecil ini, jenis data *imbalanced* akan memiliki tingkat ketepatan yang sangat rendah dalam mengklasifikasikan kategori yang sedikit dalam hal ini rumah tangga yang sebenarnya miskin ke dalam kategori miskin.

Dampak dari kesalahan klasifikasi ini, akan mempengaruhi kebijakan pemerintah dalam hal bantuan yang diberikan kepada penduduk miskin, sehingga bantuan yang diberikan tidak tepat sasaran. Dengan adanya teknik statistik dalam mengatasi data yang *imbalanced*, model yang dihasilkan meningkatkan nilai AUC menjadi sekitar 68 persen hingga 76 persen. Di sisi lain nilai *specificity* meningkat berkisar antara 67 persen sampai 78 persen. Selain itu terjadi penurunan *sensitivity* menjadi berkisar antara 68 persen sampai 70 persen. Dengan kata lain, adanya penanganan terkait data *imbalanced* membuat rata-rata nilai *specificity* dan *sensitivity* menjadi lebih berimbang yang berakibat pada rata-rata nilai akurasi secara keseluruhan menjadi lebih rendah, yaitu berkisar antara 68 persen sampai 70 persen, dibandingkan klasifikasi dengan data *imbalanced* tanpa *treatment*. Di sisi lain, kesalahan klasifikasi menjadi lebih berimbang.

Dengan mempertimbangkan rata-rata berbagai kriteria yaitu nilai *sensitivity*, *specificity*, AUC dan Kappa maka model terbaik adalah model KNN dengan skema *combine/both sampling*. Model KNN ini memiliki nilai koefisien Kappa terbesar di bandingkan model lainnya. Lebih rinci terkait performa model KNN ini memili nilai a*ccuracy* 0.73 persen, *sensitivity* sebesar 0.68 persen, s*pecificity* sebesar 78 persen, dan AUC 0.73. Nilai *accuracy* model ini sudah dikatakan baik karena sudah di atas nilai *cut off* (50 %) yang pada umumnya digunakan. Hal ini berarti bahwa model sudah mampu mengklasifikasikan dengan tepat rumah tangga sebesar 73 persen. Nilai *sensitivity* dan *specificity* cukup berimbang. Nilai AUC yang sudah di atas 0.7 juga menyatakan bahwa modelnya sudah cukup baik dalam pengklasifikasian.

4. KESIMPULAN

Adapun kesimpulan yang dapat diambil dari penelitian ini adalah bahwa model dengan menggunakan teknik *resample* menghasilkan klasifikasi yang lebih baik daripada menggunakan data asli (*imbalanced data*). Selain itu, dari keempat model klasifikasi yang diuji, model KNN memberikan performa yang terbaik jika dilihat dari aspek nilai *sensitivity*, *specificity*, AUC dan Kappa.

DAFTAR PUSTAKA

[1] E. Fissuh and M. Harris, Modeling Determinants of Poverty in Eritrea: A New Approach, pp. 1-35, 2005.

[2] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techhiques Third Edition, Waltham: Elsevier Inc, 2012.

[3] F.J. Kaunang, "Penerapan Algoritma J48 Decision Tree Untuk Analisis Tingkat Kemiskinan di Indonesia", Cogito Smart Journal, vol 4, no 2,pp 348-357, 2018

[4] B. Sartono, "Tinjauan Terhadap Keunggulan Pohon Klasifikasi Ensemble Untuk Memperbaiki Kemampuan Prediksi Pohon Klasifikasi Tunggal," BIAStatistics, vol. 9, no. 2, pp. 33-38, 2015.

[5] C. P. P. Supriyanto, "Deteksi Penyakit Diabetes Type II dengan Naive Bayes Berbasis Particle Swarm Optimization," Jurnal Teknologi Informasi, vol. 9, no. 2, 2013.

[6] H.Annur,"Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes", ILKOM Jurnal Ilmiah, vol. 10, no.2, pp 160-165, 2018

[7] W. Yustanti, "Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah," Jurnal Matematika, Statistika, & Komputasi, vol. 9, no. 1, pp. 57-68, 2012.

[8] F. Kurnia, "Klasifikasi Keluarga Miskin Menggunakan Metode K-Nearest Neighbor Berbasis Euclidean Distance", Seminar Nasional Teknologi Informasi, Komunikasi dan Industri (SNTIKI) 11, Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim Riau, pp 230-239, 2019

[9] M. Maalouf and T. Trafalis, "Rare Events and Imbalanced Datasets: An Overview," Int. Journal Data Mining, Modelling and Management, vol. 3, no. 4, pp. 375-385, 2011.

[10] C. Czado and T. Santner, "The effect of link misspecification on binary regression inference," J. Statist. Plann. Inference 33, p. 213–231. MR1190622, 1992.

[11] G. King and L. Zeng, "Logistic Regression in Rare Events Data," Journal of Political Analysis, vol. 9, no. 2, pp. 137-163, 2001.

[12] T.Purwa, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017)", JMSK, vol. 16, no. 1, pp 58-73, 2019

[13] J. Han and M. Kamber, "Data Mining Concept and Technique, Morgan: Kaufmann, 2011.

[14] J. Rodriguez, L. Kuncheva and C. Alonso, "RotationForest: A New Classifier Ensemble Method," IEEE Transactions on Pattern Analysis and Machine, vol. 28, no. 10, p. 1619–1630, 2006.

[15] Han, Jiawei, M. Kamber and J. Pei, Data Mining: Concepts and Techniques 3rd Edition, Massachusetts: Elsevier Inc, 2012.

[16] T. Fawcett, "An Introduction to ROC Analysis. Journal of Pattern Recognition Letters," An Introduction to ROC Analysis. Journal of Pattern Recognition Letters, vol. 27, pp. 861-874, 2006.

[17] F. Gorunescu, Data Mining Concept, Models and Techniques., Verlag Berlin Heidelberg: Springer, 2011.

[18] J. Landis and G. Koch, "The Measurment of Observer Agreement for Categorical Data," 2013. [Online]. Available: www.ncbi.nlm.nih.gov/pubmed/843571.