J M S K
Jurnal Matematika, Statistika & Komputasi
http:/journal.unhas.ac.id/index.php/jmsk
e-ISSN: 2614-8811
Vol.16, No. 1, 58-73, July 2019
doi: 10.20956/jmsk.v16i1.6494

Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data *Imbalanced*
(Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017)

Taly Purwa*

Abstract

This study aim is to get the best classification model for imbalanced data, i.e sample households of National Socio-Economic Survey (Susenas) March 2017 in Karangasem Regency, into the poor or not poor category. The methods used are Logistic Regression and Random Forest where for each method will be applied cross-validation (CV) scheme, i.e stratified 5-fold CV, performs under sampling, oversampling and combine sampling scheme to overcome imbalanced data problems and feature selection processes. The results showed that the application of the under sampling, oversampling and combine sampling scheme in the Logistic Regression had an effect on increasing the average value of sensitivity and decreasing average values of accuracy and specificity. However, in the Random Forest model, the effect has only appeared from the results of the under sampling scheme. The feature selection process can reduce of the variance accuracy, specificity, sensitivity and AUC in the Logistic Regression and Random Forest models only in certain schemes. The best overall model is the Logistic Regression model with a combine sampling scheme and without a feature selection process with an average value of accuracy, specificity, sensitivity and AUC of 78.13%, 79.16%, 64.44%, and 77,77% respectively.

Keywords: Poverty, Imbalanced data, Logistic Regression, Random Forest, Stratified 5-fold CV, Under sampling, Oversampling, Combine sampling

Abstrak

Penelitian ini bertujuan untuk mendapatkan model terbaik untuk klasifikasi data *imbalanced*, yaitu rumah tangga sampel Susenas Maret 2017 di Kabupaten Karangasem, ke dalam kategori miskin atau tidak. Metode yang digunakan adalah Regresi Logistik dan Random Forest dimana masing-masing diterapkan skema *cross validation* (CV), yaitu *stratified 5-fold CV*, skema *under sampling*, *oversampling* dan *combine sampling* untuk mengatasi masalah data *imbalanced* serta proses *feature selection*. Hasil penelitian menunjukkan bahwa penerapan skema *under sampling*, *oversampling* dan *combine sampling* pada model Regresi Logistik memberikan efek meningkatnya rata-rata nilai *sensitivity* dan turunnya rata-rata nilai akurasi dan *specificity*. Sedangkan pada model Random Forest, efek tersebut hanya terlihat dari hasil skema *under sampling* saja. Proses *feature selection* dapat menurunkan varian nilai akurasi, *specificity*, *sensitivity* dan AUC pada model Regresi Logistik dan Random Forest hanya pada skema tertentu. Model terbaik secara keseluruhan adalah model model Regresi Logistik dengan skema *combine sampling* dan tanpa proses *feature selection* dengan rata-rata nilai

*Badan Pusat Statistik (BPS) Provinsi Bali,
***Email**: taly@bps.go.id*

akurasi, *specificity*, *sensitivity* dan AUC masing-masing sebesar 78,13%, 79,16%, 64,44% dan 77,77%.

Kata kunci: Kemiskinan, *Imbalanced data*, Regresi Logistik, Random Forest, *Stratified 5-fold CV*, *Under sampling*, *Oversampling*, *Combine sampling*

1. Pendahuluan

Karangasem merupakan salah satu kabupaten dengan proporsi penduduk miskin tertinggi di Provinsi Bali pada tahun 2011-2017, bahkan lebih tinggi dibandingkan angka kemiskinan propinsi. Menurut data Badan Pusat Statistik (BPS) Provinsi Bali, pada tahun 2011 proporsi penduduk miskin di Kabupaten Karangasem tercatat sebesar 6.43% kemudian mengalami fluktuasi hingga mencapai angka 6.55% pada tahun 2017. Selama periode tersebut, angka terendah adalah 5.63% pada tahun 2012 sedangkan angka tertinggi adalah 7.44% pada tahun 2015.

Upaya dalam rangka menekan angka kemiskinan terus dilakukan pemerintah Provinsi Bali melalui berbagai program yang langsung menyasar ke penduduk atau rumah tangga miskin, antara lain program Jaminan Kesehatan Bali Mandara (JKBM), Gerakan Pembangunan Desa Terpadu (Gerbang Sadu), bedah rumah dan Jaminan Kredit Bali Mandara (Jamkrida). Akan tetapi data proporsi penduduk miskin yang dihasilkan oleh BPS hanya bersifat makro, atau hanya mencerminkan kondisi kemiskinan pada suatu wilayah. Oleh karena itu dibutuhkan suatu model klasifikasi yang dapat digunakan untuk mengidentifikasi rumah tangga ke dalam kategori miskin atau tidak secara akurat sehingga upaya pengentasan kemiskinan dapat tepat sasaran.

Pada tahun 2000, BPS telah melakukan Studi Penentuan Kriteria Penduduk Miskin (SPKPM) dengan menggunakan Regresi Logistik dengan tingkat akurasi sekitar 83% dan diperoleh 8 variabel yang layak sebagai penentu penduduk atau rumah tangga miskin, yaitu luas lantai per kapita, jenis lantai, ketersediaan air bersih, jenis jamban, kepemilikan asset, pendapatan per bulan, pengeluaran makanan dan konsumsi lauk pauk [1]. Akan tetapi penggunaan metode Regresi Logistik tidak tepat apabila digunakan pada kondisi data *imbalanced* (tidak seimbang), dimana model klasifikasi cenderung menihilkan peluang dari kelompok minoritas sehingga hasil prediksi akan cenderung kepada kategori mayoritas [5]. Dengan kata lain, rumah tangga yang sebenarnya miskin (kategori minoritas) cenderung akan diklasifikasikan ke rumah tangga tidak miskin (kategori mayoritas) oleh model Regresi Logistik.

Oleh karena itu, pada penelitian ini akan dilakukan perbandingan antara metode Regresi Logistik dan Random Forest untuk mengklasifikasikan rumah tangga sampel Susenas Maret 2017 di Kabupaten Karangasem, Bali. Skema yang digunakan adalah menggunakan *stratified 5-fold cross validation* (CV) pada dataset, melakukan *undersampling*, *oversampling* serta kombinasi *undersampling* dan *oversampling* sekaligus (*combine sampling*) pada data training, dan melakukan *feature selection* pada data training. Dengan skema tersebut diharapkan dapat diketahui skema mana yang dapat menghasilkan model terbaik untuk klasifikasi data kemiskinan yang bersifat *imbalance*.

2. Landasan Teori

2.1 Skema Mengatasi Imbalanced Data

Skema yang digunakan untuk mengatasi data *imbalanced* adalah dengan membuat jumlah observasi pada kedua ketegori menjadi seimbang dengan cara, yaitu *undersampling* (*without replacement*) kategori mayoritas, *oversampling* (*with replacement*) kategori minoritas dan kombinasi *undersampling* dan *oversampling* seperti yang dijelaskan pada [6].

2.2 Regresi Logistik

Regresi logistik adalah suatu model regresi yang menunjukkan pengaruh variabel prediktor, baik berupa kontinyu maupun kategorik, teradap variabel respon berupa data kategorik. Pada Regresi Logistik biner, variabel respon terdiri dari dua kategori (biner), yaitu 0 dan 1. Dimana variabel respon tersebut untuk setiap observasi mengikuti distribusi Bernoulli dan dengan model Regresi Logistik sebagai berikut,

$$\pi(x) = \frac{exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)}{1 + exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)} \tag{1}$$

Karena model diatas tidak linier pada parameter maka akan dilakukan tranformasi logit untuk mempermudah proses estimasi parameternya [4], yaitu:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{2}$$

Metode yang digunakan adalah metode *Maximum Likelihood Estimation* (MLE), dimana estimator $\boldsymbol{\beta}$ adalah nilai yang memaksimalkan fungsi log likelihood berikut,

$$L(\boldsymbol{\beta}) = ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i ln[\pi(x_i)] + (1 - \mathrm{y_i}) ln[1 - \pi(x_i)]\} \tag{3}$$

dimana $y_i = 0{,}1$ dan $i = 1,2,\dots,n$. Estimator $\boldsymbol{\beta}$ diperoleh dengan cara mendiferensialkan persamaan diatas masing-masing terhadap $\beta_0, \beta_1, \dots,$ dan β_p kemudian disamadengankan dengan nol. Karena pada persamaan tersebut mengandung $\pi(x_i)$ yang tidak linier pada parameter maka proses mendapatkan estimator harus diselesaikan dengan menggunakan metode iterasi Newton Rhapson.

Pengujian parameter secara serentak digunakan untuk mengetahui apakah variabel-variabel prediktor dalam model secara bersama-sama (serentak) berpengaruh signifikan terhadap variabel respon. Hipotesis yang digunakan dalam pengujian ini,

H_0 : $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1 : minimal ada satu $\beta_j \neq 0$, dimana $j = 1,2,\dots,p$

dengan statistik uji :

$$G = -2ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n} \hat{\pi}(x_i)^{y_i}\left(1 - \hat{\pi}(x_i)\right)^{1-y_i}}\right] \sim \chi_p^2 \tag{4}$$

dengan $n = n_0 + n_1$ dan derajat bebas p adalah jumlah variabel prediktor. Keputusan tolak H_0 jika nilai statistik uji $G > \chi^2_{(\alpha,\mathrm{p})}$ yang berarti minimal ada satu variabel prediktor yang berpengaruh signifikan terhadap variabel respon.

Pengujian parameter secara parsial untuk mengetahui apakah suatu variabel prediktor berpengaruh signifikan terhadap variabel respon. Hipotesis yang digunakan dalam pengujian ini,

H_0 : $\beta_j = 0$

H_1 : $\beta_j \neq 0$, dimana $j = 1,2,\dots,p$

dengan statistik uji Wald:

$$W = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim N(0,1) \quad (5)$$

Keputusan tolak H_0 jika nilai statistik uji $|W| > Z_{\alpha/2}$yang berarti bahwa variabel prediktor ke-j berpengaruh signifikan terhadap variabel respon.

Proses *feature selection* dilakukan dengan *backward elimination*, yaitu dengan mengeluarkan variabel dengan nilai *P-value* > α sehingga diperoleh model dengan variabel prediktor yang semuanya signifikan berpengaruh terhadap variabel respon.

2.3 Random Forest

Random forest merupakan sebuah model *ensemble*, yaitu model yang dibentuk dari banyak model Decision Tree, baik untuk regresi maupun untuk klasifikasi, dengan menerapkan metode *bootstrap aggregating* (*bagging*) dan random *feature selection*. Tahapan Random Forest adalah sebagai berikut [2]:

1. Menentukan jumlah *tree* (k) yang akan dibentuk.
2. Pengambilan sampel acak sebanyak N observasi (*with replacement*) pada dataset yang berukuran N untuk setiap *tree*.
3. Pada setiap *tree*, dilakukan juga pengambilan subset prediktor sebanyak m secara acak. Dimana $m<p$, dengan p adalah jumlah variabel prediktor.
4. Ulangi proses ke-2 dan ke-3 sampai sampai sebanyak k *tree*.
5. Pada kasus prediksi, hasil prediksi Random Forest merupakan nilai rata-rata prediksi dari sebanyak k *tree*. Sedangkan pada kasus klasifikasi, hasil prediksi Random Forest diperoleh dari vote terbanyak (*majority vote*) dari hasil klasifikasi sebanyak k *tree*.

2.4 Pengukuran Performa Model Klasifikasi

Pengukuran performa model klasifikasi dengan dua kategori dapat dilihat dari *confusion matrix* dibawah ini.

Tabel 1. *Confusion Matrix* untuk Klasifikasi dengan Dua Kategori (Biner)

Prediksi	**Aktual**		
	Positif	**Negatif**	**Total**
Positif	TP	FP	TP+FP
Negatif	FN	TN	FN+TN
Total	TP+FN	FP+TN	TP+FP+FN+TN

dengan,

TP : *True Positive*, sebenarnya positif diklasifikasikan positif
TN : *True Negative*, sebenarnya negatif diklasifikan negatif
FP : *False Positive*, sebenarnya negatif diklasifikasikan positif
FN : *False Negative*, sebenarnya positif diklasifikasikan negatif

Terdapat beberapa indikator yang dihasilkan dari *confusion matrix* diatas untuk mengukur performa klasifikasi, yaitu : pertama, *accuracy* = $\frac{\text{TP+TN}}{\text{TP+FP+FN+TN}}$, adalah indikator yang sering digunakan untuk mengukur performa klasifikasi suatu model. Akan tetapi penggunaan indikator ini kurang tepat jika diterapkan pada kasus data *imbalanced*. Kedua, *sensitivity* = $\frac{TP}{TP+FN}$,

adalah persentase observasi yang diklasifikasikan positif dari total observasi yang sebenarnya positif.

Ketiga, *specificity* = $\frac{TN}{FP+TN}$, adalah persentase observasi yang diklasifikasikan negatif dari total observasi yang sebenarnya negatif.

Selain itu terdapat indikator lain yang paling sering digunakan untuk mengukur performa klasifikasi pada data *imbalanced* adalah kurva *Receiver Operating Characteristics* (ROC) [8]. Kurva ROC menunjukkan hubungan antara 1-*specificity* dengan *sensitivity* dimana performa klasifikasi diukur dari luas area dibawah kurva ROC atau *Area Under Curve* (AUC). Luas area dibawah kurva tersebut berkisar antara 0 sampai 1. Sehingga semakin tinggi atau semakain mendekati angka 1 maka performa klasifikasi semakin baik.

3. Metode Penelitian

3.1 Sumber Data dan Variabel Penelitian

Data yang digunakan dalam penelitian ini adalah *raw* data rumah tangga sampel Susenas di Kabupaten Karangasem Bulan Maret 2017 dengan jumlah 640 rumah tangga yang diperoleh dari BPS Provinsi Bali. Dimana sebanyak 595 rumah tangga masuk kategori tidak miskin (92,97%) dan sisanya hanya sebanyak 45 rumah tangga saja yang masuk ketegori miskin (7,03%). Dengan demikian data kemiskinan tersebut merupakan data *imbalanced*. Variabel respon dan prediktor yang digunakan adalah sebagai berikut:

Tabel 2. Daftar Variabel Respon dan Variabel Prediktor

Var.	Nama	Keterangan
y	Miskin	0: Tidak Miskin 1: Miskin
x_1	Banyaknya ART	….. orang
x_2	JK KRT	0: Laki-laki 1: Perempuan
x_3	Umur KRT	0: usia produktif (15-64 th) 1: usia tidak produktif (<15 atau >64th)
x_4	Literasi KRT	0: Ya 1: Tidak
x_5	Ijasah Tertinggi KRT	0: Perguruan Tinggi 1: SMA sederajat 2: SMP sederajat 3: SD sederajat kebawah
x_6	Lapangan Usaha KRT	0: sektor non pertanian 1: sektor pertanian 2: Tidak bekerja atau menerima pendapatan
x_7	Luas lantai per kapita	….. m^2
x_8	Jenis atap terluas	0: Beton, Genteng 1: Asbes, Seng 2: Lainnya
x_9	Jenis dinding terluas	0: Tembok 1: Lainnya

Var.	Nama	Keterangan
x_{10}	Jenis lantai terluas	0: Marmer, Keramik, Parket, Ubin 1: Semen/bata merah 2: Kayu/papan, Bambu, Tanah, Lainnya
x_{11}	Fasilitas Buang Air Besar	0: Ada sendiri 1: Ada bersama 2: Ada umum atau tidak ada
x_{12}	Sumber air minum utama	0: Air kemasan bermerk, air isi ulang 1: Leding 2: Lainnya
x_{13}	Sumber penerangan utama	0: Listrik PLN dengan meteran 1: Listrik PLN tanpa meteran 2: Listrik non PLN dan bukan listrik
x_{14}	Jenis bahan bakar utama	0: Tidak memasak di rumah atau memasak dengan listrik/elpiji 12 kg/ minyak tanah 1: Elpiji 3 kg 2: Kayu Bakar
x_{15}	Tabung gas ≥5.5kg	0: Punya 1: Tidak Punya
x_{16}	Kulkas	0: Punya 1: Tidak Punya
x_{17}	Komputer/Laptop	0: Punya 1: Tidak Punya
x_{18}	Emas≥10gr	0: Punya 1: Tidak Punya
x_{19}	Sepeda Motor	0: Punya 1: Tidak Punya
x_{20}	Mobil	0: Punya 1: Tidak Punya
x_{21}	TV flat ≥30inch	0: Punya 1: Tidak Punya
x_{22}	Tanah/lahan	0: Punya 1: Tidak Punya

3.2 Tahapan Analisis

Pertama, penyiapan data penelitian, yaitu mengklasifikasikan setiap rumah tangga ke kategori miskin atau tidak dengan cara membandingkan nilai pengeluaran per kapita per bulan dengan Garis Kemiskinan (GK) Kabupaten Karangasem tahun 2017, yaitu sebesar Rp. 301.720,-. Jika pengeluaran per kapita per bulan suatu rumah tangga dibawah GK maka rumah tangga tersebut masuk kategori miskin. Selanjutnya dilakukan pengkodean ulang (*recode*) pada variabel prediktor yang bertipe kategorik sesuai Tabel 2.

Selanjutnya akan dilakukan analisis dengan menggunakan metode Regresi Logistik dan Random Forest dengan langkah berikut:

1. *Split* dataset menjadi data *training* dan *testing* (partisi 80%:20%) dengan skema *stratified 5-fold CV* sehingga terbentuk 5 set data *training* dan data *testing* berbeda.

2. Dari setiap data *training* tersebut masing-masing akan digunakan secara langsung tanpa di-*treatment* dan juga di-*treatment* dengan skema *undersampling*, *oversampling* dan *combine sampling* untuk mengatasi data *imbalanced*.
3. Pada metode Random Forest terlebih dahulu dilakukan *tuning* pada parameter *mtry*, yaitu proses penentuan jumlah subset variabel optimal yang dapat menghasilkan kesalahan klasifikasi paling kecil.
4. Menerapkan metode Regresi Logistik dan Random Forest pada setiap data *training* yang dihasilkan dari proses sebelumnya masing-masing dengan tanpa melakukan proses *feature selection* dan juga dengan proses *feature selection* pada masing-masing metode. *Feature selection* yang diterapkan pada Regresi Logistik adalah *backward elimination* variabel-variabel yang memiliki nilai P-value lebih besar dari $\alpha = 0{,}10$. Sedangkan pada Random Forest menggunakan pemilihan variabel berdasarkan *importance variable* dengan nilai threshold tertentu [10]. Pada penelitian ini digunakan nilai *threshold* sebesar 2. Dimana variabel-variabel hasil dari langkah ke-3 diatas yang memiliki nilai *importance variable* kurang dari *threshold* sebesar 2 akan dikeluarkan dari model.
5. Dari langkah ke-1 sampai ke-4 diatas akan dihasilkan sebanyak 80 hasil klasifikasi beserta indikator performa klasifikasinya, yaitu nilai *accuracy*, *sensitifity*, *specificity* dan AUC..
6. Membandingkan performa klasifikasi, dengan penilaian utama dilihat dari nilai *sensitifity*, *specificity* dan AUC, untuk mendapatkan model terbaik dari masing-masing metode.
7. Menentukan model terbaik secara keseluruhan (*overall best model*) beserta skema yang digunakan berdasarkan hasil pada tahap sebelumnya.

Pada penelitian ini, nilai *sensitivity* menunjukkan tingkat ketepatan dalam mengklasifikasikan rumah tangga yang sebenarnya miskin ke dalam ketegori miskin juga. Sedangkan nilai *specificity* menunjukkan tingkat ketepatan dalam mengklasifikasikan rumah tangga yang sebenarnya tidak miskin ke kategori tidak miskin juga. Resiko yang diakibatkan dari kesalahan mengklasifikasikan rumah tangga miskin ke kategori tidak miskin (1- *sensitifity*) dan kesalahan mengklasifikasikan rumah tangga tidak miskin ke kategori miskin (1- *specificity*) dianggap sama besarnya.

4. Hasil dan Pembahasan

4.1 Klasifikasi dengan Regresi Logistik

Klasifikasi data *imbalanced* tanpa *treatment* dengan model Regresi Logistik secara umum menghasilkan nilai rata-rata akurasi dan *specificity* yang lebih tinggi dan nilai rata-rata *sensitivity* yang jauh lebih rendah dibandingkan data yang sudah di-*treatment* untuk mengatasi *imbalance* dengan skema *undersampling*, *oversampling* serta *combine sampling*. Sedangkan untuk rata-rata nilai AUC, kedua skema tersebut menghasilkan nilai yang relatif tidak berbeda, yaitu berkisar antara 58,96% sampai 77,77%.

Nilai rata-rata akurasi dan *specificity* tertinggi yang mampu dihasilkan oleh data *imbalanced* tanpa *treatment* masing-masing sebesar 93,13% dan 100,00%. Sedangkan rata-rata nilai *sensitivity* yang dihasilkan hanya sebesar 11,11% untuk model tanpa *feature selection* dan 2,22% untuk model dengan *feature selection*. Artinya klasifikasi dengan data *imbalanced* memiliki tingkat ketepatan yang sangat rendah dalam mengklasifikasikan rumah tangga yang sebenarnya miskin ke dalam kategori miskin. Akibatnya akan sangat banyak rumah tangga miskin yang tidak tersentuh oleh program pengentasan kemiskinan dari pemerintah.

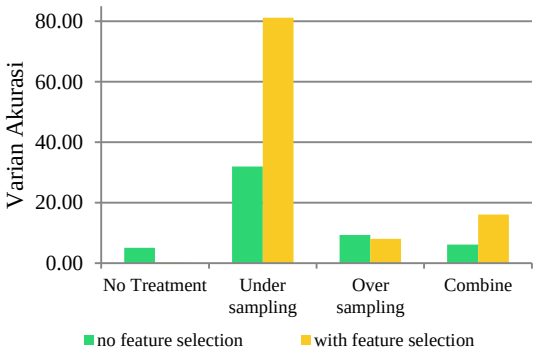
Penerapan skema untuk mengatasi data *imbalanced*, baik dengan *under sampling*, *oversampling* maupun *combine sampling*, mampu meningkatkan rata-rata nilai *sensitivity* sampai berkisar antara 60,00% sampai 70,37% sekaligus membuat rata-rata nilai specificity mengalami

penurunan menjadi berkisar antara 55,63% sampai 79,16%. Dengan kata lain, skema tersebut 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 55,94% sampai 78,13%, dibandingkan klasifikasi dengan data *imbalanced* tanpa *treatment*. Presentase kesalahan klasifikasi, baik kesalahan mengklasifikasikan rumah tangga miskin ke kategori tidak miskin (1- *sensitifity*) maupun kesalahan mengklasifikasikan rumah tangga tidak miskin ke kategori miskin (1- *specificity*), menjadi lebih berimbang.

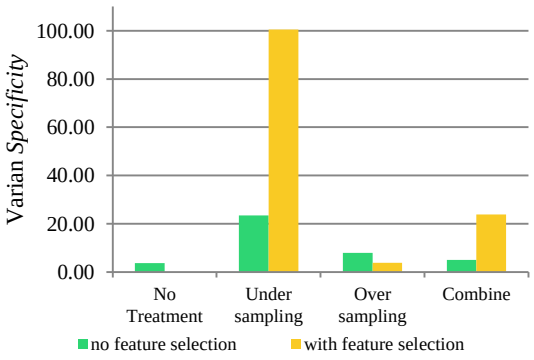
Tabel 3. Rata-rata dan Varian Performa Klasifikasi Model Regresi Logistik

Skema	*Feature Selection*	*Accuracy*	*Specificity*	*Sensitivity*	*AUC*
Tanpa *treatment*	No	92.50	98.66	11.11	77.40
		(5.07)	*(3.74)*	*(61.73)*	*(160.61)*
	Yes	93.13	100.00	2.22	74.90
		(0.12)	*(0.00)*	*(24.69)*	*(15.47)*
Undersampling	No	55.94	55.63	60.00	58.96
		(31.92)	*(23.44)*	*(345.68)*	*(168.23)*
	Yes	62.50	61.90	70.37	71.44
		(81.18)	*(100.51)*	*(164.61)*	*(32.72)*
Oversampling	No	76.41	77.65	60.00	74.96
		(9.28)	*(7.98)*	*(407.41)*	*(160.51)*
	Yes	73.28	73.95	64.44	75.23
		(8.06)	*(3.88)*	*(271.60)*	*(35.41)*
Combine sampling	No	78.13*	79.16*	64.44*	77.77*
		(6.10)	*(5.08)*	*(271.60)*	*(57.91)*
	Yes	73.75	74.62	62.22	76.26
		(16.05)	*(23.80)*	*(222.22)*	*(26.54)*

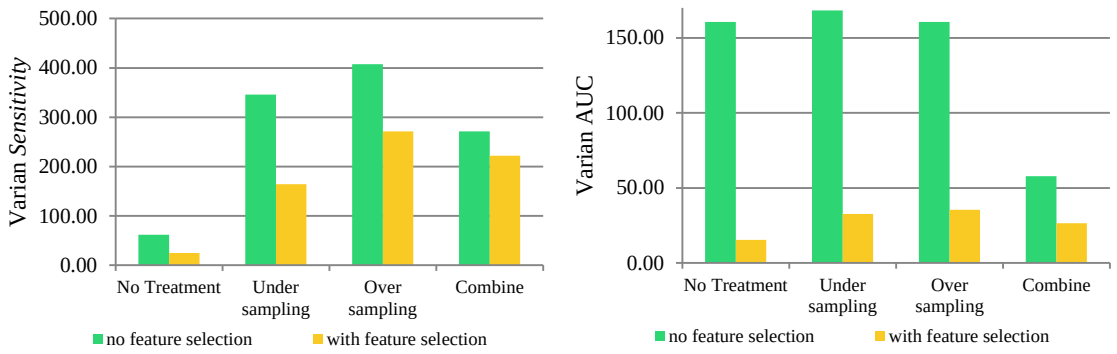
*Keterangan : varian dalam kurung ( ); *) model Regresi Logistik terbaik*



(a)



(b)



(c) (d)

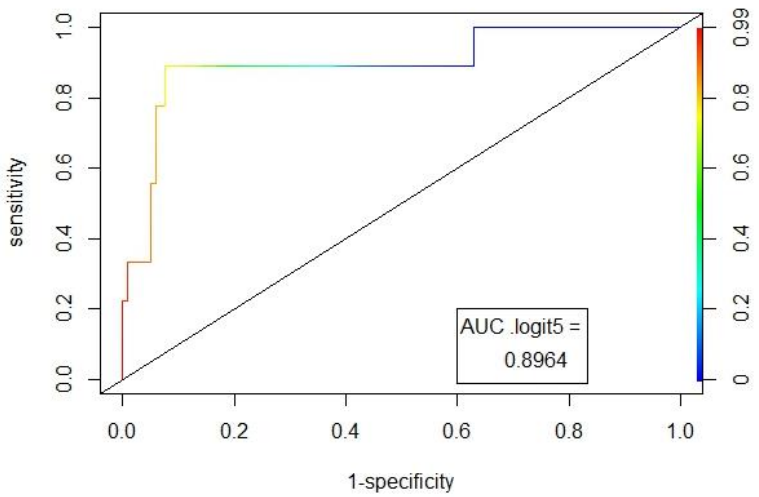
Gambar 1. Varian Nilai Akurasi (a), *Specificity* (b), *Sensitivity* (c) dan AUC (d) Berdasarkan Skema Mengatasi Data *Imbalanced* dan Proses *Feature Selection* untuk Model Regresi Logistik

Dari gambar diatas, secara umum klasifikasi dengan data yang telah di-*treatment* untuk mengatasi *imbalance* menghasilkan varian nilai akurasi, *specificity* dan *sensitivity* yang lebih besar dibandingkan dengan data tanpa *treatment*. Data dengan skema *combine sampling* menghasilkan varian yang relatif lebih kecil dibandingkan skema *undersampling* dan *oversampling*. Secara umum, proses *feature selection* pada model Regresi Logistik dapat menurunkan varian nilai akurasi, *specificity*, *sensitivity* dan AUC dibandingkan tanpa proses *feature selection*. Akan tetapi hal tersebut tidak berlaku untuk varian nilai akurasi dan *specificity* yang dihasilkan dari skema *undersampling* dan *combine sampling* pada Gambar 1 (a) dan (b), dimana proses *feature selection* justru menghasilkan varian yang jauh lebih tinggi dibandingkan tanpa proses *feature selection*.

Dengan mempertimbangkan rata-rata nilai *specificity*, *sensitivity* dan AUC maka model Regresi Logistik terbaik adalah model dengan skema *combine sampling* dan tanpa proses *feature selection*. Jika memperhatikan hasil klasifikasi lebih rinci pada setiap *fold*, seperti disajikan pada Lampiran 1, maka model Regresi Logistik terbaik dihasilkan oleh *fold* ke-5 dengan skema *oversampling* tanpa *feature selection* dengan nilai akurasi 81,25%, *sensitivity* 88,89%, *specificity* 80,67% dan AUC 89,64% dengan persamaan sebagai berikut :

$$g(x) = ln\left[\frac{P(Y=1|x)}{P(Y=0|x)}\right]$$
$$\begin{aligned} &= -133{,}44 + 0{,}18x_1 + 1{,}15x_2 + 0{,}64x_3 - 2{,}18x_4 + 15{,}07x_{51} \\ &+ 14{,}72x_{52} + 14{,}62x_{53} - 0{,}44x_{61} - 0{,}20x_{62} - 0{,}20x_7 - 0{,}10x_{81} \\ &- 23{,}13x_{82} + 0{,}19x_9 - 0{,}34x_{101} - 1{,}20x_{102} + 0{,}68x_{111} + 0{,}92x_{112} \\ &+ 16{,}18x_{121} + 15{,}27x_{122} + 0{,}20x_{131} + 1{,}47x_{132} + 14{,}91x_{141} \\ &+ 16{,}57x_{142} + 18{,}35x_{15} + 1{,}54x_{16} + 16{,}78x_{17} + 18{,}46x_{18} \\ &+ 1{,}19x_{19} + 16{,}85x_{20} + 16{,}52x_{21} + 0{,}92x_{22} \end{aligned} \quad (6)$$

dengan kurva ROC yang disajikan pada Gambar 2.



Gambar 2. Kurva ROC dan AUC Model Regresi Logistik *fold* ke-5 dengan Skema *Oversampling* tanpa *Feature Selection*

4.2 Klasifikasi dengan Random Forest

Sama dengan hasil pada model Regresi Logistik, klasifikasi dengan model Random Forest dengan data *imbalanced* tanpa *treatment* menghasilkan nilai akurasi dan *specificity* yang lebih tinggi dan nilai rata-rata *sensitivity* yang lebih rendah dibandingkan data yang sudah di-*treatment*. Akan tetapi hasil berbeda ditunjukkan dari data yang sudah di-*treatment*, khususnya dengan skema *oversampling* dan *combine sampling* yang tidak mampu meningkatkan nilai *sensitivity* secara signifikan, nilainya hanya berkisar antara 13% sampai 31% saja. Kedua skema tersebut menghasilkan nilai akurasi, *specificity*, *sensitifity* dan AUC yang relatif sama dengan klasifikasi dengan data *imbalanced* tanpa *treatment.* Rata-rata nilai AUC yang dihasilkan model Random Forest, baik dengan data *imbalanced* tanpa *treatment* maupun dengan *treatment*, memiliki besaran yang tidak terlalu bervariasi, yaitu berkisar antara 70,51% sampai 78,46%.

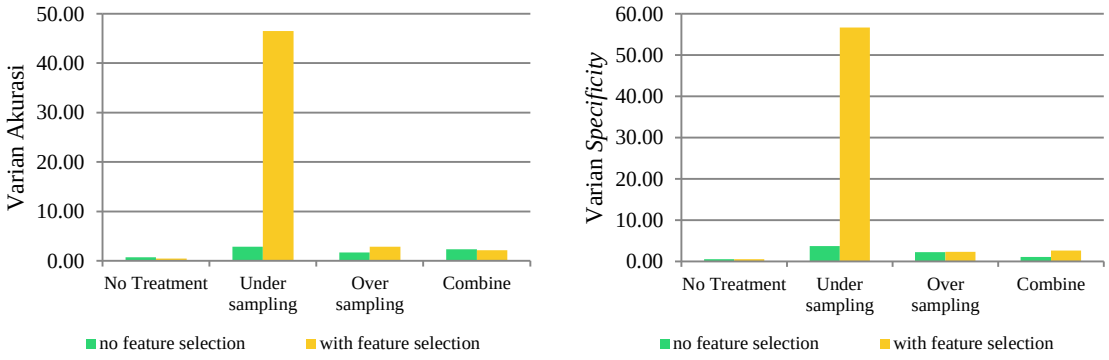
Nilai rata-rata akurasi dan *specificity* tertinggi yang mampu dihasilkan oleh data *imbalanced* tanpa *treatment* masing-masing sebesar 92,81% dan 99,66%. Sedangkan rata-rata nilai *sensitivity* yang dihasilkan hanya sebesar 2,22% untuk model tanpa *feature selection* maupun dengan *feature selection*. Dampak *treatment* untuk mengatasi data *imbalanced* hanya terlihat dari skema *undersampling* dimana nilai *sensitivity* dapat meningkat hingga mencapai 80,00% tanpa proses *feature selection* dan 66,67% dengan proses *feature selection*. Sedangkan rata-rata nilai akurasi dan *specificity* mengalami penurunan dibandingkan dengan klasifikasi dengan data tanpa *treatment* dengan nilai masing-masing sebesar 63,91% dan 62,69% tanpa proses *feature selection* serta 65,00% dan 64,87% dengan proses *feature selection*.

Tabel 4. Rata-rata dan Varian Performa Klasifikasi Model Random Forest

Skema	*Feature Selection*	*Accuracy*	*Specificity*	*Sensitivity*	*AUC*
Tanpa *treatment*	No	92.81	99.66	2.22	78.46
		(0.73)	*(0.56)*	*(24.69)*	*(68.35)*
	Yes	92.66	99.50	2.22	76.18
		(0.49)	*(0.56)*	*(24.69)*	*(60.61)*
Undersampling	No	63.91*	62.69*	80.00*	76.36*
		(2.87)	*(3.74)*	*(86.42)*	*(28.17)*

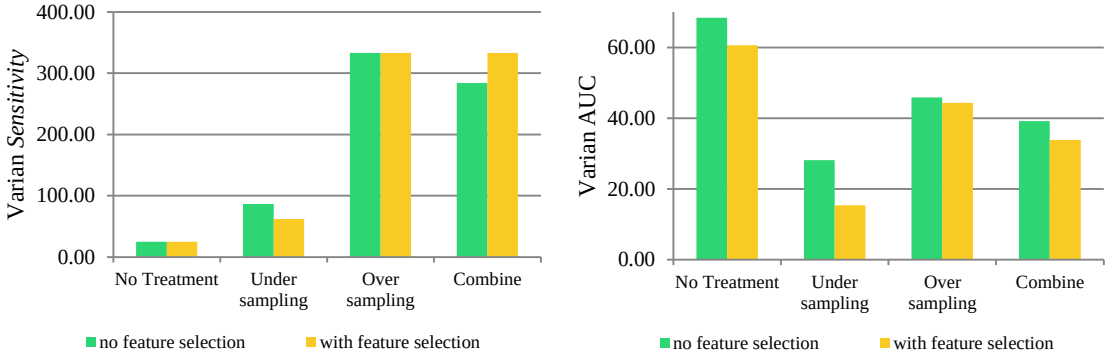
Skema	*Feature Selection*	*Accuracy*	*Specificity*	*Sensitivity*	*AUC*
	Yes	65.00	64.87	66.67	70.51
		(46.51)	*(56.63)*	*(61.73)*	*(15.37)*
Oversampling	No	90.16	95.97	13.33	75.97
		(1.71)	*(2.26)*	*(333.33)*	*(45.87)*
	Yes	90.47	96.30	13.33	77.92
		(2.87)	*(2.33)*	*(333.33)*	*(44.36)*
Combine sampling	No	87.19	91.60	28.89	76.07
		(2.32)	*(1.06)*	*(283.95)*	*(39.18)*
	Yes	86.72	90.92	31.11	76.34
		(2.14)	*(2.61)*	*(333.33)*	*(33.81)*

*Keterangan : varian dalam kurung ( ); *) model Random Forest terbaik*



(a) (b)

Gambar 3. Varian Nilai Akurasi (a), *Specificity* (b), *Sensitivity* (c) dan AUC (d) Berdasarkan Skema Data *Imbalanced* dan Proses *Feature Selection* untuk Model Random Forest



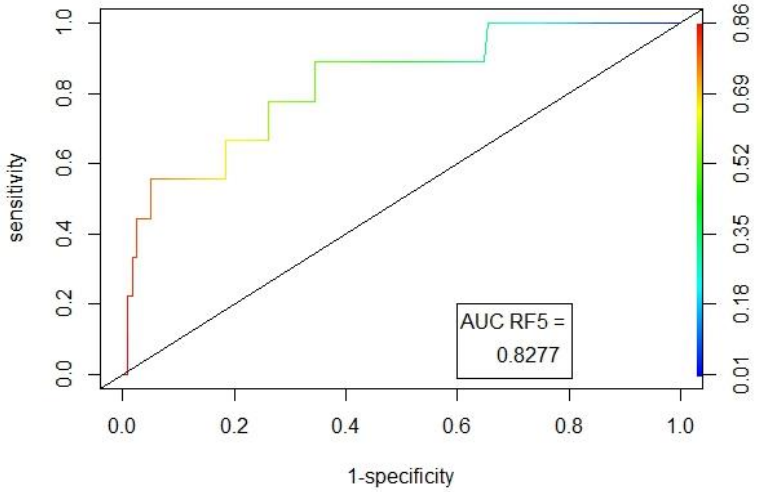
(c) (d)

Gambar 3. (Lanjutan)

Sama dengan hasil model Regresi Logistik, klasifikasi dengan model Random Forest pada data yang telah di-*treatment* untuk mengatasi *imbalance* menghasilkan varian nilai akurasi, *specificity* dan *sensitivity* yang lebih besar dibandingkan dengan data tanpa *treatment*. Sebaliknya,

varian nilai AUC dari data tanpa *treatment* lebih besar dibandingkan dengan varian nilai AUC dari data yang sudah di-*treatment* seperti terlihat pada Gambar 3 (d). Proses *feature selection* pada model Random Forest hanya dapat menurunkan varian untuk nilai AUC saja (Gambar 3 (d)). Sedangkan skema *undersampling* dengan proses *feature selection* justru menghasilkan varian nilai akurasi dan *specificity* yang jauh lebih besar daripada skema yang sama tanpa proses *feature selection* seperti disajikan pada Gambar 3 (a) dan (b).

Dengan mempertimbangkan kriteria yang sama dengan model Regresi Logistik maka model Random Forest terbaik adalah model dengan skema *undersampling* dan tanpa proses *feature selection*. Jika memperhatikan hasil klasifikasi lebih rinci pada setiap *fold*, seperti disajikan pada Lampiran 2, maka model Random Forest terbaik dihasilkan oleh *fold* ke-5 dengan skema *undersampling* tanpa *feature selection* dengan nilai akurasi 64,84%, *specificity* 63,03%, *sensitivity* 88,89% dan AUC 82,77% dengan nilai parameter *mtry* hasil *tuning* sebesar 4, jumlah *tree* sebanyak 500 dan kurva ROC yang disajikan pada Gambar 4.



Gambar 4. Kurva ROC dan AUC Model Random Forest *fold* ke-5 dengan Skema *Undersampling* tanpa *Feature Selection*

4.3 Penentuan Model Terbaik Secara Keseluruhan (*Overall Best Model*)

Perbandingan performa klasifikasi model terbaik dari model Regresi Logistik dan Random Forest disajikan pada Tabel 5. Kedua model terbaik memiliki nilai AUC yang tidak jauh berbeda, yaitu sebesar 77,77% untuk model Regresi Logistik dan 76,36% untuk model Random Forest. Model Regresi Logistik dengan skema *combine sampling* tanpa proses *feature selection* memiliki rata-rata nilai *specificity* dan AUC yang lebih besar dibandingkan metode Random Forest dengan skema *undersampling* tanpa proses *feature selection*. Sedangkan model Random Forest dengan skema *undersampling* tanpa proses *feature selection* memiliki rata-rata nilai *sensitivity* yang lebih besar dibandingkan Regresi Logistik dengan skema *combine sampling* tanpa proses *feature selection*. Berdasarkan hasil tersebut maka model klasifikasi terbaik secara keseluruhan adalah model Regresi Logistik dengan skema *combine sampling* tanpa proses *feature selection*.

Tabel 5. Rata-rata dan Varian dari Model Terbaik Regresi Logistik dan Random Forest

Metode	Skema	*Accuracy*	*Specificity*	*Sensitivity*	*AUC*
Regresi Logistik	- *Combine sampling*	78.13*	79.16*	64.44*	77.77*

	- No feature selection	*(6.10)*	*(5.08)*	*(271.60)*	*(57.91)*
Random Forest	*- Undersampling*	63.91	62.69	80.00	76.36
	- No feature selection	*(2.87)*	*(3.74)*	*(86.42)*	*(28.17)*

*Keterangan : varian dalam kurung ( ); *) overall best model*

5. Kesimpulan dan Saran

Berdasarkan hasil dan pembahasan diatas dapat diperoleh beberapa kesimpulan dari penelitian ini, yaitu:

1. Penerapan skema *undersampling*, *oversampling* dan *combine sampling* untuk menangani masalah data *imbalanced* pada model Regresi Logistik mampu menghasilkan rata-rata nilai *sensitivity* yang lebih besar dibandingkan tanpa menerapkan skema tersebut. Sehingga resiko yang diakibatkan dari kesalahan mengklasifikasikan rumah tangga miskin ke kategori tidak miskin berupa tidak tersentuhnya rumah tangga miskin oleh program pengentasan kemiskinan dari pemerintah dapat diminimalisir. Akan tetapi untuk model Random Forest hanya skema *undersampling* saja yang dapat menghasilkan rata-rata nilai *sensitivity* yang lebih besar. Selain itu efek penerapan skema *undersampling*, *oversampling* dan *combine sampling* adalah berkurangnya rata-rata nilai akurasi dan *specificity* dibandingkan tanpa menerapkan skema tersebut. Sedangkan rata-rata nilai AUC yang dihasilkan, baik dengan menerapkan skema maupun tidak, relatif tidak berbeda.
2. Proses *feature selection* dapat menurunkan varian nilai akurasi, *specificity*, *sensitivity* dan AUC pada model Regresi Logistik dan Random Forest, kecuali untuk beberapa skema tertentu.
3. Berdasarkan rata-rata nilai *specificity*, *sensitivity* dan AUC maka model Regresi Logistik terbaik adalah model dengan skema *combine sampling* dan tanpa proses *feature selection*. Sedangkan model Random Forest terbaik adalah model dengan skema *undersampling* dan tanpa proses *feature selection*.
4. Model klasifikasi terbaik secara keseluruhan adalah model Regresi Logistik dengan skema *combine sampling* dan tanpa proses *feature selection*.

Saran yang dapat diberikan untuk penelitian selanjutnya adalah perlunya penggunaan *treatment* untuk mengatasi data *imbalanced* (*sampling based-method*) lainnya, seperti *Tomek Link* (T-Link) oleh [9] dan *Synthetic Minority Oversampling Technique* (SMOTE) oleh [3]. Selain itu, penggunaan pengembangan metode Regresi Logistik yang dapat mengatasi data *imbalanced*, yaitu *Rare Event Weighted Logistic Regression* (RE-WLR) oleh [7], perlu juga dipertimbangkan. Pada metode Random Forest, perlu dilakukan *tuning* terhadap parameter number of *tree* sehingga diperoleh jumlah yang optimal untuk meningkatkan akurasi model Random Forest.

Daftar Pustaka

[1] BPS., 2016. *Perhitungan dan Analisis Kemiskinan Makro Indonesia 2016*. Jakarta: Badan Pusat Statistik.

[2] Breiman, L., 2001. Random Forest. *Machine Learning*, Vol. 45, No. 1, hal. 5-32.

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

[4] Hosmer, D.W. dan Lemeshow, S. 2000. *Applied Logistic Regression: second edition*. New Jersey : John Wiley & Sons, Inc.

[5] King, G dan Zeng, L., 2001. Logistic Regression in Rare Events Data. *Political Analysis*, Vol. 9, No. 2, hal. 137-163.

[6] Lunardon, N., Menardi, G. dan Torelli, N., 2014. ROSE: A Package for Binary Imbalanced Learning. *The R Journal*, Vol. 6, No. 1, hal. 79-89.

[7] Maalouf, M. dan Siddiqi, M., 2014. Weighted Logistic Regression for LargeScale Imbalanced and Rare Events Data, *Journal of Knowledge Based Systems*, Vol. 59, hal. 142-148.

[8] Menardi, G. dan Torelli, N., 2012. Training and Assessing Classification Rules with Imbalanced Data. *Data Mining Knowledge Discovery*, Vol. 28, No. 1, hal. 92-122.

[9] Tomek, I., 1997. Two Modifications of CNN. *IEEE Transactions of Systems Man and Communications*, Vol 6, No. 11, hal. 769-772.

[10] Torgo, L., 2011. *Data Mining with R: Learning with Case Studies*. Boca Raton : Chapman & Hall/CRC press.

Lampiran 1

Hasil Lengkap Model Regresi Logistik

Skema	CV	*Feature Selection*	*Accuracy*	*Specificity*	*Sensitivity*	AUC
No Treatment	1	No	89.06	95.80	0.00	76.84
		Yes	92.97	100.00	0.00	73.11
	2	No	94.53	100.00	22.22	85.34
		Yes	93.75	100.00	11.11	75.96
	3	No	91.41	97.48	11.11	78.06
		Yes	92.97	100.00	0.00	69.47
	4	No	93.75	100.00	11.11	56.86
		Yes	92.97	100.00	0.00	75.86
	5	No	93.75	100.00	11.11	89.92
		Yes	92.97	100.00	0.00	80.11
Undersampling	1	No	51.56	52.10	44.44	48.74
		Yes	0.00	0.00	0.00	0.00
	2	No	51.56	52.10	44.44	48.74
		Yes	52.34	50.42	77.78	65.13
	3	No	62.50	60.50	88.89	78.34
		Yes	65.63	66.39	55.56	72.92
	4	No	52.34	52.10	55.56	52.85

Skema	CV	*Feature Selection*	*Accuracy*	*Specificity*	*Sensitivity*	AUC
		Yes	0.00	0.00	0.00	0.00
	5	No	61.72	61.34	66.67	66.11
		Yes	69.53	68.91	77.78	76.28
Oversampling	1	No	75.00	76.47	55.56	73.86
		Yes	75.00	75.63	66.67	71.99
	2	No	75.00	75.63	66.67	82.07
		Yes	71.88	72.27	66.67	77.36
	3	No	73.44	74.79	55.56	73.58
		Yes	71.88	73.11	55.56	75.16
	4	No	77.34	80.67	33.33	55.65
		Yes	70.31	72.27	44.44	67.88
	5	No	81.25	80.67	88.89	89.64
		Yes	77.34	76.47	88.89	83.75
Combine sampling	1	No	76.56	78.15	55.56	75.91
		Yes	76.56	78.99	44.44	78.15
	2	No	75.78	76.47	66.67	81.70
		Yes	67.97	67.23	77.78	72.92
	3	No	77.34	78.15	66.67	76.19
		Yes	76.56	78.15	55.56	77.78
	4	No	78.91	81.51	44.44	67.27
		Yes	71.09	72.27	55.56	69.56
	5	No	82.03	81.51	88.89	87.77
		Yes	76.56	76.47	77.78	82.91

Lampiran 2

Hasil Lengkap Model Random Forest

Skema	CV	*Feature Selection*	*Accuracy*	*Specificity*	*Sensitivity*	AUC
No Treatment	1	No	91.41	98.32	0.00	67.41
		Yes	91.41	98.32	0.00	64.61
	2	No	93.75	100.00	11.11	80.02
		Yes	92.97	99.16	11.11	75.91
	3	No	92.97	100.00	0.00	76.98
		Yes	92.97	100.00	0.00	77.50
	4	No	92.97	100.00	0.00	77.36
		Yes	92.97	100.00	0.00	76.38
	5	No	92.97	100.00	0.00	90.52
		Yes	92.97	100.00	0.00	86.51
Undersampling	1	No	62.50	61.34	77.78	70.03
		Yes	71.09	71.43	66.67	67.32
	2	No	66.41	65.55	77.78	78.34
		Yes	53.91	52.94	66.67	66.90
	3	No	62.50	60.50	88.89	78.90
		Yes	68.75	69.75	55.56	68.91
	4	No	63.28	63.03	66.67	71.76

Skema	CV	*Feature Selection*	*Accuracy*	*Specificity*	*Sensitivity*	AUC
		Yes	67.97	68.07	66.67	75.07
	5	No	64.84	63.03	88.89	82.77
		Yes	63.28	62.18	77.78	74.37
Oversampling	1	No	89.06	95.80	0.00	68.21
		Yes	89.06	95.80	0.00	70.68
	2	No	92.19	98.32	11.11	81.09
		Yes	92.19	98.32	11.11	82.49
	3	No	89.06	95.80	0.00	76.19
		Yes	88.28	94.96	0.00	76.56
	4	No	89.84	95.80	11.11	70.31
		Yes	91.41	97.48	11.11	73.06
	5	No	90.63	94.12	44.44	84.03
		Yes	91.41	94.96	44.44	86.79
Combine sampling	1	No	86.72	92.44	11.11	68.49
		Yes	87.50	93.28	11.11	68.86
	2	No	86.72	91.60	22.22	80.25
		Yes	86.72	91.60	22.22	80.72
	3	No	88.28	92.44	33.33	73.76
		Yes	86.72	89.92	44.44	73.95
	4	No	85.16	89.92	22.22	73.44
		Yes	84.38	89.08	22.22	74.65
	5	No	89.06	91.60	55.56	84.41
		Yes	88.28	90.76	55.56	83.52