Statistika, Vol. 21 No. 1, 11 – 20
Mei 2021

Klasifikasi Status Bekerja Individu di Provinsi Banten Tahun 2020 dengan Menggunakan Metode LASSO dan Adaptive LASSO

PARDOMUAN ROBINSON SIHOMBING[1], KHAIRIL A. NOTODIPUTRO[2], BAGUS SARTONO[3]

[1]Badan Pusat Statistik (BPS), Jakarta, Indonesia
[1,2,3]Department Statistika, IPB University, Bogor, Indonesia
e-mail: robinson@bps.go.id

ABSTRAK

Penelitian ini bertujuan membandingkan metode LASSO dan Adaptive LASSO dengan penggunaan imbalanced data pada regresi binary logistik. Studi kasus yang digunakan adalah pemodelan klasifikasi status bekerja individu di Provinsi Banten tahun 2020. Hasil yang didapat performa LASSO maupun Adaptive LASSO memberikan hasil yang sama baiknya. Dengan mempertimbangkan berbagai kriteria performa dalam accuracy, sensitivity dan *specificity*, maka model terbaik adalah model LASSO dengan simulasi data balanced 60 persen dan 40 persen dengan nilai masing-masing sebesar 79,16 persen; 80.29 persen dan 68,75 persen. Terdapat beberapa paradoks/anomali dalam hasil penelitian di antaranya peluang status tidak bekerja seseorang menurut lokasi tempat tinggal, gender dan pendidikan. Status disabilitas masih menjadi masalah dalam mencari pekerjaan. Semakin banyak anggota rumah tangga maka akan semakin tinggi peluangnya berstatus tidak bekerja. Semakin tinggi usia seseorang maka akan semakin kecil peluangnya berstatus tidak bekerja. Peluang status tidak bekerja seseorang yang menikah lebih kecil daripada yang belum/tidak kawin.

Kata Kunci: *Adaptive*, *Imbalanced*, Lasso, Logistik, Regresi.

ABSTRACT

This study aims to compare the LASSO and Adaptive LASSO methods with the use of unbalanced data in binary logistic regression. The case study used is modeling the classification of individual employment status in Banten Province in 2020. The results obtained by the performance of LASSO and Adaptive LASSO give the same good results. By considering various performance criteria in terms of accuracy, sensitivity and specificity, the best model is the LASSO model with a balanced simulation of 60 percent and 40 percent of data with a value of 79.16 percent, respectively; 80.29 percent and 68.75 percent. There are several paradoxes/anomalies in the research results, including the probability of a person's unemployment status based on location of residence, gender and education. Disability status is still an obstacle in finding work. The more household members, the higher the chance of not working. The older a person is, the less likely they are to become unemployed. The chance of not working for someone who is married is smaller than that of a single/unmarried person.

Keywords: Adaptive, Imbalance, Lasso, Logistic, Regression.

1. PENDAHULUAN

Salah satu masalah dalam pembangunan ekonomi adalah Tingkat Pengangguran Terbuka (TPT) selain Kemiskinan, Distribusi Pendapatan dan Pertumbuhan Ekonomi. Menurut BPS, TPT adalah persentase jumlah pengangguran terhadap jumlah angkatan kerja. Berbagai usaha telah dilakukan pemerintah untuk menekan angka TPT. Berdasarkan data dari BPS, angka TPT Indonesia terus menurun dari tahun 2011-2019 yaitu sebesar 7,14 tahun di tahun 2011 menjadi 5,23 persen di tahun 2019. Provinsi Banten adalah provinsi yang selalu memiliki TPT tertinggi di Indonesia, walaupun merupakan daerah yang memiliki tingkat konsentrasi yang

tinggi untuk industri manufaktur. Data TPT Provinsi Banten berfluktuasi setiap tahun selalu di atas 8 persen. Pada tahun 2020, angka TPT Banten meningkat dari sebelumnya 8,11 persen di tahun 2019 menjadi lebih dari 10 persen. Hal ini salah satunya sebagai dampak dari adanya pandemi virus Corona yang melanda Indonesia maupun dunia. Tingginya angka TPT ini ditengah prediksi Indonesia yang akan mencapai bonus demografi di tahun 2030, akan menjadi suatu masalah dan beban bagi negara, sehingga penting untuk mengetahui faktor apa saja yang mempengaruhi status bekerja seseorang sehingga dapat diberikan stimulus maupun kebijakan yang tepat sasaran.

Prediksi status bekerja individu dapat dilakukan dengan pemodelan klasifikasi. Klasifikasi status bekerja individu dapat dibagi menjadi bekerja dan tidak bekerja (menganggur). Klasifikasi status bekerja individu dapat dilakukan salah satunya dengan menggunakan model regresi logistik, Ada banyak variabel yang dapat mempengaruhi status bekerja seseorang dari dari sisi internal berupa gender, umur, pendidikan, keahlian dan ada tidaknya disabilitas yang diderita, maupun dari sisi eksternal berupa status perkawinan, lokasi tempat tinggal dan banyaknya anggota rumah tangga. Karena banyaknya variabel yang mempengaruhi status bekerja ini, terdapat kemungkinan antar variabel penjelas/kovariat yang memiliki korelasi. Selain itu semakin banyak variabel penjelas akan menyulitkan dalam intrepetasi sehingga dibutuhkan suatu teknik dalam menyeleksi variabel penjelas.

Salah satu metode dalam menyeleksi variabel adalah metode LASSO (Model Least Absolute Shrinkage and Selection Operator) yang dikembangkan oleh Tibshirani (1996). Beberapa penelitian menggunakan metode LASSO di antaranya Putranto (2017) menggunakan metode Regresi Logistik dengan Teknik LASSO, Stepwise dan Komponen Utama dalam Klasifikasi Data Curah Hujan, hasil yang didapat bahwa regresi logistik LASSO menghasilkan ketepatan dan spesifisitas paling tinggi untuk menduga data curah hujan dibandingkan metode komponen utama dan metode stepwise. Metode LASSO dapat mengatasi keterbatasan ketika ada banyak prediktor yang dianalisis, dengan cara menyusutkan variabel dengan perkiraan yang sangat tidak stabil menuju nol, model LASSO dapat secara efektif mengecualikan beberapa variabel yang tidak relevan. Namun, dalam praktiknya, Fan dan Li (2001) menyatakan bahwa LASSO dapat menciptakan bias yang berlebihan ketika memilih variabel yang signifikan dan tidak konsisten dalam hal pemilihan variabel. Di sisi lain menurut Fan dan Li (2001), model LASSO tidak memenuhi kondisi "*oracle property*". Hal ini diperkuat oleh Meinshausen dan Bühlmann (2006) menunjukkan bahwa pemilihan variabel dengan LASSO dapat konsisten jika mendasari model memenuhi beberapa kondisi. Zou (2006) mengembangkan model Adaptive LASSO untuk mengatasi kelemahahan yang ada pada model LASSO biasa, dimana sudah memenuhi kondisi "oracle property" dan model ini menambakan penimbang dalam model LASSO. Penelitain terkait adaptive LASSO juga dilakukan oleh Ardiansyah, dkk (2020) dengan melihat peningkatan presisi dugaan berat gabah melalui proses seleksi peubah dalam pembelajaran mesin statistika.

Masalah lain yang sering dihadapi dalam model klasifikasi seperti Regresi logistik adalah modelnya didasarkan pada asumsi bahwa banyaknya data terdistribusi secara merata antara kelas yang berbeda. Menurut Maalouf dan Trafalis (2011) umumnya justru data yang dikumpulkan proporsinya tidak sama yang disebut dengan imbalanced data. Apabila data yang digunakan adalah 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, hal ini terutama data yang digunakan adalah data dalam jumlah yang besar (*big data*) (King & Zeng, 2001).

Berdasarkan permasalahan yang telah disebutkan di atas, pada penelitian ini akan mengkaji dan menerapkan metode regresi logistik dengan membandingkan performa metode LASSO dan adaptive LASSO. Selain itu menerapkan metode imbalanced data dalam pemodelan klasifikasi status bekerja di Provinsi Banten.

2. METODE PENELITIAN

Informasikan Data yang digunakan dalam penelitian ini berasal dari Survei Angkatan Kerja Nasional (Sakernas) Provinsi Banten periode Agustus 2020, yang dilakukan oleh Badan Pusat Statistik (2021). Adapun total sampel yang digunakan adalah 11.469 responden, dimana 10,2 persen berstatus tidak bekerja, sisanya 89,8 persen berstatus bekerja. Adapun variabel yang digunakan dalam penelitian dapat dilihat pada Tabel 1.

Tabel 1 Variabel Penelitian

Nama Variabel	Keterangan	Skala
Status Bekerja	0 Bekerja, 1 Tidak Bekerja	Nominal
Tipe Daerah	0 Perkotaan, 1 Perdesaan	Nominal
Jenis Kelamin	0 Wanita, 1 Pria	Nominal
Status Perkawinan	0 Belum Kawin, 1 Kawin, 2 Cerai	Nominal
Pendidikan	0 Tidak Sekolah,1 SD, 2 SMP, 3 SMA, 4 Perguruan Tinggi(PT)	Ordinal
Kursus	0 Tidak, 1 Ya	Nominal
Dis_lihat	0 Tidak, 1 Ya	Nominal
Dis_dengar	0 Tidak, 1 Ya	Nominal
Dis_jalan	0 Tidak, 1 Ya	Nominal
Dis_pegang	0 Tidak, 1 Ya	Nominal
Dis_bicara	0 Tidak, 1 Ya	Nominal
Dis_lainnya	0 Tidak, 1 Ya	Nominal
Jml_art		Rasio
Umur		Rasio

Regresi Logistik

Regresi logistik merupakan salah salah analisis regresi yang digunakan apabila variabel respon berupa data kategorik. Apabila variabel responnya hanya terdiri dari dua kategori (*biner*), yaitu 0 dan 1 maka dikenal dengan regresi *binary logistik*. Variabel respon yang digunakan akan mengikuti distribusi Bernoulli untuk setiap observasi dan dengan model Regresi Logistik yang digunakan sebagai berikut:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad \text{..... (1)}$$

LASSO

Metode LASSO dikembangkan pertama kali oleh Thibsirani (1996), dimana model yang digunakan mampu menyusutkan koefisien regresi tepat nol sehingga dapat digunakan untuk menyeleksi peubah. Penduga koefisien regresi LASSO disusutkan ke arah nol seiring dengan peningkatan nilai λ (hyper-parameter non negatif) yang digunakan. Parameter yang diduga dengan meminimalkan fungsi:

$$\widehat{\boldsymbol{\beta}}_{lasso}^{*(n)}(\text{logistic}) = \underset{\boldsymbol{\beta}}{\arg\min} \sum_{j=1}^{n}\left(-y_j \boldsymbol{x}_j'\boldsymbol{\beta} + \log\left(1+\exp\left(\boldsymbol{x}_j'\boldsymbol{\beta}\right)\right)\right) + \lambda_n \sum_{j=1}^{p}|\beta_j| \quad \text{..... (2)}$$

Adaptive LASSO

Adaptive LASSO merupakan pengembangan dari model LASSO dengan menggunakan bobot adaptive untuk mengatasi kelemahan pada model LASSO (Zou, 2006). Pendugaan parameter pada adaptive Lasso yaitu dengan meminimumkan fungsi:

$$\widehat{\boldsymbol{\beta}}_{alasso}^{*(n)}(\textbf{logistic}) = \underset{\boldsymbol{\beta}}{\textbf{arg min}} \sum_{j=1}^{n}\left(-y_j \boldsymbol{x}_j'\boldsymbol{\beta} + \textbf{log}\left(\textbf{1}+\textbf{exp}\left(\boldsymbol{x}_j'\boldsymbol{\beta}\right)\right)\right) + \lambda_n \sum_{j=1}^{p}\hat{w}_j|\beta_j| \quad \text{.....} \quad (3)$$

dengan $\widehat{w}_j = \frac{1}{|\hat{\beta}_j|^r}$ dimana r adalah bilangan positif dan $\widehat{\boldsymbol{\beta}}_j$ diduga dengan regresi ridge yaitu dengan meminimumkan fungsi:

$$\widehat{\boldsymbol{\beta}}_{ridge}^{*(n)}(\text{logistic}) = \underset{\boldsymbol{\beta}}{\arg\min} \sum_{j=1}^{n}\left(-y_j \boldsymbol{x}_j'\boldsymbol{\beta} + \log\left(1+\exp\left(\boldsymbol{x}_j'\boldsymbol{\beta}\right)\right)\right) + \lambda_n \sum_{j=1}^{p}\beta_j^2 \quad \text{..... (4)}$$

Modeling dan Resampling

Tahap modeling data dibagi ke dalam 2 kelompok yaitu data training data dan testing data dengan proporsi 70 persen untuk data training yang digunakan untuk membangun model sementara 30 persen untuk data testing yang akan digunakan untuk menghitung performance dari model yang terbentuk dengan membandingkan label data sebenarnya dan label data hasil

klasifikasi model. Selanjutnya dari data training, dilakukan resample dengan teknik both sampling dengan lima kondisi untuk masing-masing proporsi kategori status bekerja dan tidak bekerja sebesar 80 persen dan 20 persen, 70 persen dan 30 persen, 60 persen dan 40 persen, 55 persen dan 45 persen serta 50 persen dan 50 persen.

K-Fold Cross-Validation (K-Fold CV)

Menurut Nisbet, dkk (2009), *cross-validation* merupakan bentuk dari resampling yang mengambil beberapa sampel dari keseluruhan observasi dan menjadikannya sebagai *training* data untuk model. Dalam penelitian ini menggunakan 10-Fold Cross-Validation untuk mendapatkan nilai lamda (λ) yang meminimumkan fungsi CV pada model LASSO, Ridge dan adaptive LASSO. Pola yang digunakan dalam 10-Fold Cross-Validation, menggunakan 9-fold sebagai training dan 1 fold sebagai testing secara bergantian.

Evaluasi

Dalam mengevaluasi model predictive dilakukan dengan mengetahui sejauh mana pengklasifikasian dapat mengenal atau memprediksi kelas data yang dapat dilihat dari Confussion Matrix. Menurut Han, dkk (2012), *confussion matrix* merupakan tabel berukuran mxm dengan m=jumlah kelas. Bagian kolom diisi oleh label aktual untuk tiap kelas, sementara bagian baris diisi oleh label kelas hasil prediksi.

Tabel 2 *Confusion Matrix*

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

Ukuran akurasi digunakan untuk melihat kebaikan suatu model. Akurasi merupakan proporsi frekuensi yang tepat diklasifikasikan dengan total sampel yang ada. Selain melihat akurasi perlu melihat sensitivity dan specificity. Sensitivity merupakan proporsi kelas mayor yang menjadi perhatian/diinginkan terprediksi dengan benar. Specificity merupakan proporsi kelas lainnya/minor yang juga menjadi perhatian terprediksi dengan benar. Han (2012) menyatakan seharusnya ketiga ukuran tersebut bernilai besar dan seimbang, apabila tingkat akurasi tinggi, namun sensitivity dan specificity rendah, maka pengklasifikasian dapat dikatakan tidak baik.

$$Akurasi = \frac{TP+TN}{P+N} \quad (5)$$

$$Sensitivity+= \frac{TP}{P} \quad (6)$$

$$Specificity-= \frac{TN}{N} \quad (7)$$

3. HASIL DAN PEMBAHASAN

Hasil Langkah awal di dalam menganalisis data adalah menampilkan statistik deskriptif pada variabel penelitian. Statistik deskriptif digunakan untuk melihat gambaran awal dari data sampel yang digunakan. Dalam analisis deskriptif ini menggunakan tabulasi silang antara variabel dependen (status bekerja) dengan variabel independennya. Salah satu faktor yang mempengaruhi status bekerja adalah faktor internal dari individu. Faktor internal yang dimaksud adalah faktor yang melekat pada individu tersebut tidak berkaitan dengan individu lainnya. Pada Tabel 3. menampilkan proporsi status bekerja pada variabel gender, pendidikan, umur dan kepemilikan sertifkat kursus. Jika dilihat dari jenis gender, proporsi perempuan yang tidak bekerja lebih sedikit dari laki-laki yang tidak bekerja. Hal ini dapat diartikan bahwa dibandingkan lelaki, perempuan memiliki kecenderungan memiliki kemudahan untuk mendapatkan pekerjaan. Selanjutnya adalah proporsi tingkat pendidikan terhadap status bekerja. Dari data yang ada terlihat proporsi yang tidak bekerja justru tertinggi pada level Pendidikan SMP dan SMA. Hal ini dapat diakibatkan lulusan SMP dan SMA akan melanjutkan studinya ke jenjang yang lebih tinggi. Jika dilihat dari sisi umur, status tidak bekerja terbanyak pada rentang usia 15-30 tahun. Usia ini kebanyakan masih akan melanjutkan studinya dan sedang mencari pekerjaan karena baru menyelesaikan pendidikannya. Berbeda pada rentang usia selanjutnya yang pada umumnya masuk pada usia mapan atau sudah

bekerja. Kepemilikan sertifikat kelulusan tidak terlalu berdampak terhadap status pengangguran di Banten.

Tabel 3 Proporsi Status Bekerja dengan Faktor Internal Individu (persen) Provinsi Banten, 2020

Variabel	Indikator	Bekerja	Tidak Bekerja	Total
Gender	Laki-laki	89,2	10,8	100
	Perempuan	91,0	9,0	100
Pendidikan	Tidak Tamat SD	94,5	5,5	100
	SD Sederajat	94,6	5,4	100
	SMP Sederajat	88,6	11,4	100
	SMA Sederajat	84,4	15,6	100
	PT	93,3	6,7	100
Umur	15-30	75,2	24,8	100
	31-55	95,2	4,8	100
	>55	95,3	4,7	100
Sertifkat Kursus	Tidak	89,9	10,1	100
	Ya	89,6	10,4	100

Faktor selanjutnya yang dapat mempengaruhi status bekerja seseorang adanya faktor eksternal, diman faktor ini berkaitan dengan interaksi individu lainnya. Pada Tabel 4 menampilkan proporsi status bekerja pada variabel daerah tempat tinggal, jumlah anggota rumah tangga (ART) dan status perkawinan. Jika dilihat dari daerah tempat tinggal, proporsi daerah perkotaan yang tidak bekerja lebih banyak daripada daerah perdesaan. Hal ini berkaitan dengan kompetisi dan kompetensi daerah perkotaan yang cenderung lebih tinggi daripada daerah perdesaan. Selanjutnya adalah proporsi jumlah ART terhadap status bekerja. Dari data yang ada terlihat proporsi yang tidak bekerja berbanding lurus dengan jumlah ART. Dalam sebuah rumah tangga yang berukuran besar akan memiliki kecendungan anggota rumah tangganya belum memiliki pekerjaan yang akan menjadi tanggungan Kepala Rumah Tangga (KRT). Selanjutnya adalah status perkawinan seseorang. Proporsi status tidak bekerja tertinggi berada pada status belum/tidak kawin. Pada level ini ada kemungkinan individu tersebut masih menjadi tanggungan dari Kepala Rumah Tangga, sedangkan pada kelompok dengan proporsi status tidak bekerja terendah berada pada status kawin.

Tabel 4 Proporsi Status Bekerja dengan FaktorEkternal Individu (persen) Provinsi Banten, 2020

Variabel	Indikator	Bekerja	Tidak Bekerja	Total
daerah	perdesaan	90,40	9,60	100
	perkotaann	89,52	10,48	100
Jumlah ART	1-2	93,34	6,66	100
	3-5	89,90	10,10	100
	>5	85,26	14,74	100
Status Perkawinan	Belum/ Tidak Kawin	71,85	28,15	100
	Kawin	95,82	4,18	100
	Cerai	92,97	7,03	100

Selanjutnya adalah melakukan analisis inferensia terhadap hubungan variabel dependen dengan variabel independennya. Dalam penelitian ini menggunakan model regresi binary logistik karena variabel dependennya berupa data kategori. Dalam model yang digunakan ditambahkan penalty dengan dua pendekatan yaitu pendekatan LASSO dan Adaptive LASSO. Pada Tabel 5 dapat dilihat performa pendekatan LASSO dengan beberapa simulasi data. Jika dilihat dari sisi accuracy dan sensitivity maka model dengan data asli (data imbalanced) memberikan performa terbaik. Dengan nilai masing-masing sebesar 90,24 persen dan 9,.96

persen. Akan tetapi nilai specificity sangat kecil hanya sebesar 0,29 persen. Nilai specificity menunjukan performa model dalam memprediksi frekuensi pada kelas yang minor (sedikit) sehingga ada kecenderungan akan adalah kesalahan klasifikasi pada kelas yang minor. Selanjutnya dilakukan simulasi dengan membalanced data dengan menggunakan teknik resample dengan 5 skenario dengan perbandingan kelas mayor dan minor masing-masing menjadi 80:20, 70:30, 60:0, 55:45 dan 50:50. Pada Tabel 5, terlihat semakin data mendekati data balanced maka nilai specificity terus meningkat, akan tetapi sebagai trade off nilai accuracy dan sensitivity terus menurun, Nilai specificity tertinggi pada simulasi data ke-5 sebesar 69,35 persen akan tetapi memiliki accuracy dan sensitivity terendah dengan besaran masing-masing 77,56 persen dan 78,35 persen. Di sisi lain koefisien Kappa berfluktuasi, dimana nilainya meningkatkan sampai pada simulasi kedua dan menurun dari simulasi ke-3 hingga ke-5. Jumlah parameter yang nyata juga bervariasi, yang terbesar pada simulasi ke-3 sebanyak 17 parameter.

Tabel 5 Performa Model LASSO

LASSO	data asli	simulasi1	simulasi2	simulasi3	simulasi4	simulasi5
	90:10	80:20	70:30	60:40	55:45	50:50
lamda	0,0016	0,0015	0,0013	0,0010	0,0013	0,0017
Accuracy	90,24	87,1	87,13	79,16	78,06	77,56
Sensitivity	99,96	92,14	92,01	80,29	79,07	78,45
Specificity	0,29	40,48	41,96	68,75	68,75	69,35
Kappa	0,0048	0,3082	0,3174	0,2932	0,2775	0,2730
Jumlah parameter yang nyata	13	13	13	17	16	16

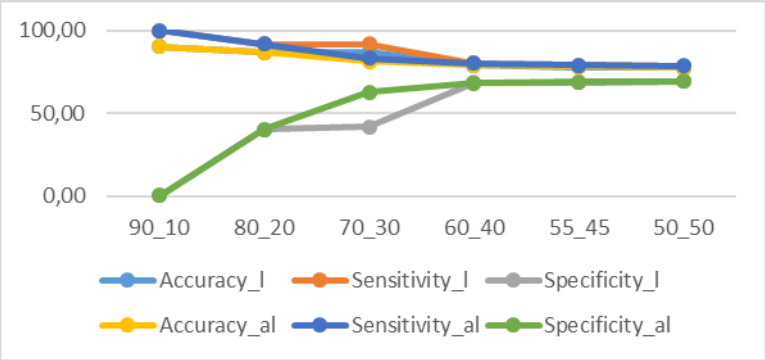
Pembahasan selanjutnya adanya performa model Adaptive LASSO Model ini merupakan pengembangan dari model LASSO. Pada Tabel 6, sama seperti model LASSO, terlihat dari sisi accuracy dan sensitivity maka model dengan data asli (data imbalanced) memberikan performa terbaik dengan nilai masing-masing sebesar 90,24 persen dan 99,97 persen. Senada dengan model LASSO, nilai specificity model Adaptive LASSO dengan data asli juga sangat kecil sebesar 0,3 persen. Selanjutnya dari hasil simulasi dengan membalanced data, hasilnya menunjukan semakin mendekati data balanced maka nilai specificity terus meningkat, akan tetapi sebagai trade off nilai accuracy dan sensitivity terus menurun, Nilai specificity tertinggi pada simulasi data ke-5 sebesar 69,35 persen akan tetapi memiliki accuracy dan sensitivity terendah dengan besaran masing-masing 77,71 persen dan 78,62 persen. Di sisi lain koefisien Kappa berfluktuasi, dimana nilainya meningkatkan sampai pada simulasi kedua dan menurun dari simulasi ke-3 hingga ke-5. Jumlah parameter yang nyata juga bervariasi, yang terbesar pada simulasi data ke-5 sebanyak 17 parameter dan terendah pada simulasi data ke-2 sebanyak 11 parameter.

Tabel 6 Performa Model Adaptive LASSO

Adaptive LASSO	Asli	simulasi1	simulasi2	simulasi3	simulasi4	simulasi5
	90:10	80:20	70:30	60:40	55:45	50:50
lamda ridge	0,0090	0,0160	0,0202	0,0226	0,0231	0,0234
lamba LASSO	0,0019	0,0010	0,0091	0,0017	0,0007	0,0013
Accuracy	90,24	86,98	81,40	79,08	78,26	77,71
Sensitivity	99,97	92,01	83,41	80,26	79,26	78,62
Specificity	0,30	40,48	62,80	68,15	69,05	69,35
Kappa	0,004	0,3054	0,3045	0,2896	0,2814	0,2749
Jumlah parameter yang nyata	15	13	11	16	15	17

Selanjutnya jika dibandingkan antara model LASSO dan Adaptive LASSO keduanya memberikan hasil peforma yang sama baiknya pada Gambar 1. Nilai specifity yang terus meningkatkan senada dengan jumlah data menuju data balanced. Pada data asli hingga data

simulasi ke-3 performa (accuracy, sensitivity, dan specificity) model LASSO lebih baik dibandingkan dengan model adaptive LASSO, akan tetapi berlaku sebaliknya pada data simulasi ke-4 dan ke-5, model adaptive LASSO memili performa yang lebih baik.



Gambar 1 Perbandingan Performa Lasso dan Adptive Lasso

Jika membandingkan nilai rata-rata proporsi untuk ketiga kriteria (*accuracy, sensitivity*, dan *specificity)* baik secara simultan maupun masing-masing menggunakan Wilcoxon Sign Rank Test maka dapat dikatakan tidak ada perbedaan nilai antara metode lasso dan adaptive lasso. Pada Tabel 7, dapat dilihat nilai z-stat masih di bawah nilai z tabel untuk alpa lima persen sbesar 1,96. Selanjutnya dilakukan uji proporsi untuk masing-masing kriteria pada setiap simulasi dengan menggunakan uji z. Dari hasil uji dua sampel proporsi, terlihat bahwa terdapat perbedan nilai proporsi pada simulasi kedua (70:30), dengan nilai proporsi lasso lebih tinggi untuk kriteria accuracy dan sensitivity, sementara adaptive lasso lebih tinggi untuk kriteria specificity. Hal ini dapat dilihat nilai z-stat di atas nilai z tabel untuk alpa lima persen sbesar 1,96. Pada data asli dan simulasi lainnya perbedaan nilai proporsi untuk masing-masing kriteria dianggap tidak signifikan.

Tabel 7 Nilai Z-Statistik Uji Proporsi Lasso dan Adaptive Lasso

Simulasi	Accuracy	Sensitivity	Specificity
90_10	0.000	-0.211	-0.284
80_20	0.037	0.190	0.000
70_30	6.527	10.321	-5.408
60_40	0.082	0.030	0.167
55_45	-0.201	-0.184	-0.084
50_50	-0.149	-0.163	0.000
Mean	0.893	0.917	0.465
Grandmean		0.865	

Jika membandingkan berbagai kriteria yang ada dari semua model, maka model terbaik adalah model LASSO dengan simulasi data 60:40, dimana kriteria *accuracy, sensitivity*, dan *specificity* yang lebih seimbang, dengan besaran masing-masing 79,16 persen, 80,29 persen dan 68,75 persen. Selanjutnya adalah pembahasan terkait hasil estimasi parameter koefisien regresi.

Tabel 8 Koefisien Regresi Logistik

Variabel	Indikator	Koefisien	Odds
(Intercept)	(Intercept)	0,953114	2,59
daerah (perkotaan)	perdesaan	-0,06666	0,94
Jumlah ART	Jumlah ART	0,065362	1,07
Gender (Perempuan)	Laki-laki	0,200886	1,22
Umur	Umur	-0,02774	0,97
Status Perkawinan	Kawin	-1,67004	0,19
	Cerai	-0,63053	0,53
Pendidikan	SD Sederajat	,	-
	SMP Sederajat	0,227305	1,26
	SMA Sederajat	0,238501	1,27
	Perguruan Tinggi	-0,37984	0,68
Kursus (Tidak)	kursus_sertifikat	-0,05939	0,94
dis_lihat (tidak)	dis_lihat	0,080142	1,08
dis_dengar (tidak)	dis_dengar	0,800257	2,23
dis_jalan (tidak)	dis_jalan	0,106256	1,11
dis_pegang (tidak)	dis_pegang	0,223074	1,25
dis_bicara (tidak)	dis_bicara	1,18773	3,28
dis_lainnya (tidak)	dis_lainnya	0,092484	1,10

() baseline

Koefisien variabel daerah menunjukkan arah negatif, dengan *odds* sebesar 0,94 hal ini bermakna bahwa peluang status tidak bekerja seseorang yang tinggal di perdesaan lebih kecil sebesar 0,94 kali dari seseorang yang tinggal di perkotaan. Hal senada juga ditemukan oleh Dhanani (2004), dengan menggunakan data Indonesia menemukan bahwa tingkat pengangguran terbuka di kota lebih besar tiga kali lipat dibandingkan dengan di daerah perdesaan, pada rentang waktu 1976-2000. Hal ini selain dikarenakan jumlah penduduk yang tingal di perkotaan (urban) lebih banyak dibandingkan di perdesaan (rural), kompetisi dalam mencari lapangan pekerjaan juga lebih besar di daerah perkotaan. Hal ini semakin diperparah dengan adanya migrasi penduduk dari perdesaan ke perkotaan yang mengakibatkan jumlah ketersedian lapangan pekerjaan di perkotaan semakin sedikit dibandingkan jumlah penduduk yang mencari pekerjaan.

Koefisien variabel jumlah angota rumah tangga menunjukkan arah positif, hal ini bermakna bahwa semakin besar ukuran jumlah anggota rumah tangga dalam suatu keluarga maka peluang anggota rumah tangganya berstatus tidak bekerja akan meningkat. Hal senada dengan teori yang dikemukakan oleh Sumarsono (2003), yang menyatakan bahwa semakin banyak anggota dalam tiap-tiap keluarga yang mengurus rumah tangga semakin kecil tingkat partisipasi kerja yang pada akhirnya akan menambah jumlah individu yang tidak bekerja.

Koefisien variabel gender menunjukkan arah positif dengan *odds* sebesar 1,22, hal ini bermakna peluang status tidak bekerja laki-laki lebih tinggi 1,22 kali daripada perempuan. Hal senada ditemukan oleh Mutiadanu, dkk (2018) yang menyatakan peluang laki-laki menganggur lebih besar dari perempuan. Hal ini dikarenakan kecenderungan perempuan lebih mudah menyesuaikan dengan kondisi lapangan pekerjaan yang ada dan dapat melakukan pekerjaan dari rumah.

Koefisien variabel umur menunjukkan arah negatif, hal ini bermakna bahwa semakin dewasa umur seseorang maka peluang status tidak bekerjanya akan semakin kecil. Hal senada juga ditemukan oleh Dhanani (2004), dimana tingkat pengangguran terbuka pemuda beberapa kali lebih tinggi daripada orang dewasa. Hal ini berkaitan dengan kemapanan dan pengalaman yang dimiliki oleh seseorang seiring bertambahnya usia.

Koefisien variabel status perkawinan menunjukkan arah negatif dengan *odds* sebesar 0,19 untuk status kawin dan 0,53 untuk kasus cerai, hal ini bermakna bahwa peluang status tidak bekerja seseorang yang kawin mapun cerai lebih kecil dibandingkan yang belum/tidak kawin. Hal ini senada dengan Yuliatin, dkk (2011)yang menyatakan bahwa salah satu faktor yang mempengaruhi seseorang akan rajin bekerja adalah adanya tanggungan keluarga yang harus dipenuhi. Hal ini mengakibatkan peluang tidak bekerja, seseorang yang memiliki status kawin akan memiliki peluang yang lebih kecil daripada yang belum/tidak kawin.

Koefisien variabel pendidikan untuk level SD tidak signifikan terhadap yang tidak bersekolah. Koefisien untuk level SMP dan SMA menunjukkan arah positif dengan *odds* sebesar 1,26 untuk kategori SMP dan 1,27 untuk kategori SMA, hal ini bermakna bahwa peluang status tidak bekerja seseorang berpendidikan SMP dan SMA lebih tinggi masing-masing sebesar 1,26 kali dan 1,27 kali daripada yang tidak bersekolah. Hal sebaliknya untuk level Perguruan tinggi dengan keofisien bertanda positif dengan odd sebesar 0,68. Hal ini bermakna bahwa peluang status tidak bekerja seseorang Perguruan Tinggi lebih rendah sebesar 0,68 kali daripada yang tidak bersekolah. Hal senada ditemukan oleh Mutiadanu, dkk (2018) serta Dhanani (2004) yang menyatakan peluang menganggur pada tingkat pendidikan SMP dan SMA lebih tinggi dari lebih pendidikan di bawahnya.

Koefisien variabel sertifikat pendidikan menunjukkan arah negatif dengan *odds* sebesar 0,94, hal ini bermakna bahwa peluang status tidak bekerja seseorang yang memiliki sertifikat pelatihan/keterampilan lebih kecil sebesar 0,9 kali dibandingkan yang tidak memiliki sertifikat. Hal ini senada dengan penelitian Pasay dan Indrayanti (2012) yang menyatakan bahwa peluang bekerja seseorang yang memiliki sertifikat keahlian/keterampilan akan lebih tinggi daripada yang tidak memiliki. Hal ini karena dengan adanya sertifkat kursus/keahlian dianggap sebagai investasi dan peningkatan kualitas seseorang.

Koefisien variabel dissabilitas baik untuk disfungsi penglihatan, pendengaran, bicara, gerak dan lainnya menunjukkan arah positif, hal ini bermakna bahwa peluang status tidak bekerja seseorang yang memiliki disfungsi lebih tinggi daripada yang tidak memiliki disfungsi. Hal ini senada dengan penelitian Cahyono (2017) yang menyatakan bahwa kesulitan memperoleh hak dan kesempatan yang sama memdapatkan pekerjaan yang layak di sektor perusahaan swasta dan pemerintah sesuai kemampuan.

4. SIMPULAN

Berdasarkan pembahasan di atas, maka dapat disimpulkan bahwa performa LASSO maupun Adaptive LASSO memberikan hasil yang sama baiknya. Semakin data mendekati data balanced maka nilai specificity terus meningkat, akan tetapi sebagai trade off nilai accuracy dan sensitivity terus menurun. Dengan mempertimbangkan berbagai kriteria yang ada dari semua model, maka model terbaik adalah model LASSO dengan simulasi data 60:40, hal ini dengan mengkonfirmasi kestabilan performa yang dilakukan kedua model. Terdapat beberapa paradoks/anomali dalam hasil penelitian di antaranya peluang status tidak bekerja seseorang yang tinggal di perkotaan lebih besar dari yang tinggal di perdesaan, peluang status tidak bekerja laki-laki lebih besar dari perempuan, dan peluang status tidak bekerja seseorang yang memiliki Pendidikan SMP dan SMA lebih besar daripada yang tidak tamat sekolah. Peluang status tidak bekerja seseorang yang memiliki disabilitas baik untuk disabilitas penglihatan, pendengaran, bicara, gerak dan lainnya lebih tinggi daripada yang normal.

Semakin banyak anggota rumah tangga maka akan semakin tinggi peluang seseorang berstatus tidak bekerja. Semakin tinggi usia seseorang maka akan semakin kecil peluangnya berstatus tidak bekerja. Peluang status tidak bekerja seseorang yang menikah lebih kecil daripada yang belum/tidak kawin. Seseorang yang memiliki sertifikat keahlian akan cenderung memiliki peluang status tidak bekerja yang lebih kecil daripada yang tidak memiliki sertifikat.

DAFTAR PUSTAKA

Ardiansyah, M., Notodiputro, K. A., & Sartono, B. (2020). Peningkatan Presisi Dugaan Berat Gabah Melalui Proses Seleksi Peubah Dalam Pembelajaran Mesin Statistika. Prosiding Seminar Nasional VARIANSI, (pp. 171-183).

Badan Pusat Statistik. (2021). Labor Market Indicators Indonesia August 2020. Jakarta: Badan Pusat Statistik.

Cahyono, S. A. (2017). Penyandang Disabilitas: Menelisik Layanan Rehabilitasi Sosial Difabel Pada Keluarga Miskin. Media Informasi Penelitian Kesejahteraan Sosial, 41(3), 239-254.

Dhanani. (2004). Unemployment and Underemployment in Indonesia,. Switzeland: International Labour Office.

Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. Journal of the American Statistical Association, 96, 1348–1360.

Han, Jiawei, Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques 3rd Edition. Massachusetts: Elsevier Inc.

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. Journal of Political Analysis, 9(2), 137-163.

Maalouf, M., & Trafalis, T. (2011). Rare Events and Imbalanced Datasets: An Overview. . Int. Journal Data Mining, Modelling and Management, 3(4), 375-385.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. Annals of statistics, 34(3), 1436-1462.

Mutiadanu, S., Adry, M. R., & Putri, D. Z. (2018). Analisis Sosial Ekonomi Terhadap Pengangguran Muda. Ecosains, 7(2), 89-98.

Nisbet, Robert, Elder, J., & Miner, G. (2009). Handbook of Statistical Analysis and Data. California: Elsevier Inc.

Pasay, N., & Indrayanti, R. (2012). Pengangguran, Lama Mencari Kerja, dan Reservation Wage Tenaga Kerja Terdidik. Jurnal Ekonomi danPembangunan Indonesia, 12(2), 116-135.

Putranto, N., Silvianti, P., & Soleh, A. M. (2017). Klasifikasi Data Curah Hujan Menggunakan Metode Regresi Logistik dengan Teknik Lasso, Stepwise dan Komponen Utama. Skripsi, Statistika. Bogor: IPB.

Sumarsono, S. (2003). Ekonomi Manajemen Sumber Daya Manusia dan Ketenagakerjaan. Yogyakarta: Graha Ilmu.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society, 58(1), 267–288

Yuliatin, Huseno, T., & Febriani. (2011, Mei). Pengaruh Karakteristik Kependudukan Terhadap Pengangguran di Sumatera Barat. Jurnal Manajemen dan Kewirausahaan, 2(2).

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association, 101(476), 1418-1429.