JAMBURA JOURNAL OF MATHEMATICS
Jambura J. Math. Vol. 5, No. 1, pp. 166–178, February 2023
Journal Homepage: http://ejurnal.ung.ac.id/index.php/jjom
DOI: https://doi.org/10.34312/jjom.v5i1.17050

Synthetic Minority Oversampling Technique Pada Model Logit dan Probit Status Pengangguran Terdidik

Fatimah[1,*], Anwar Fitrianto[1], Indahwati[1], Erfiani[1], Khusnia Nurul Khikmah[1]

[1]*Departemen Statistika, Fakultas MIPA, IPB University, Bogor 16680, Indonesia*
[*]*Corresponding author. Email: twfatimah@apps.ipb.ac.id*

ABSTRAK

Pengangguran terdidik disebabkan karena kurang selarasnya perencanaan pembangunan pendidikan dengan perkembangan lapangan pekerjaan sehingga lulusan berbagai institusi pendidikan tidak terserap oleh lapangan kerja. Data pengangguran terdidik di DKI Jakarta terdapat kelas yang tidak seimbang. Data tidak seimbang merupakan masalah yang serius dalam pemodelan karena dapat menyebabkan terjadinya kesalahan prediksi sehingga berpengaruh terhadap akurasi model. Penggunaan *Synthetic Minority Oversampling Technique* (SMOTE) untuk penanganan data tidak seimbang diduga mampu menaikkan akurasi model. Penelitian ini bertujuan: mencari model terbaik untuk mengidentifikasi faktor-faktor yang berpengaruh terhadap status pengangguran terdidik menggunakan model logit dan probit serta melakukan penanganan data yang tidak seimbang menggunakan SMOTE. Hasil penelitian menunjukkan bahwa variabel bebas yang berpengaruh terhadap status pengangguran terdidik pada model logit dan probit sama yaitu kelompok umur dan keikutsertaan dalam pelatihan, sedang pada model logit dan probit dengan SMOTE, variabel bebas yang berpengaruh terhadap status pengangguran terdidik juga sama yaitu kelompok umur, status perkawinan, dan keikutsertaan dalam pelatihan. Penanganan data tidak seimbang dengan menggunakan SMOTE mampu menaikkan nilai *balanced accuracy* secara signifikan. Hal ini terlihat dari nilai *balanced accuracy* untuk model baik logit dan probit dengan SMOTE lebih tinggi dibanding model logit dan probit tanpa SMOTE. Model logit dengan SMOTE merupakan model terbaik karena memiliki nilai *balanced accuracy* tertinggi dibanding model lain. Berdasarkan model logit dengan SMOTE, pengangguran terdidik di DKI Jakarta berasal dari kelompok umur muda dan berstatus belum pernah kawin. Perlu adanya peran pemerintah dalam meningkatkan kualitas institusi pendidikan dalam menghasilkan lulusan yang sesuai kualifikasi perusahaan, sehingga mampu terserap oleh perusahaan penyedia lapangan kerja. Pengangguran yang pernah mengikuti pelatihan walaupun berpendidikan tinggi ternyata juga berpotensi untuk menjadi pengangguran. Bekal pelatihan yang dimiliki ternyata belum mampu menekan angka pengangguran, olehkarena itu pemerintah hendaknya mampu menyediakan pelatihan-pelatihan yang dapat meningkatkan kemampuan berwirausaha dan sekaligus menyediakan modal berupa kredit usaha sehingga dapat menekan angka pengangguran terdidik.

Kata Kunci:
SMOTE; Logit; Probit; Pengangguran Terdidik

ABSTRACT

Educated unemployment is caused by a misalignment of educational development planning and employment development, resulting in underemployed graduates from various educational institutions.

Received: 23 November 2022 | Accepted: 25 January 2023 | Online: 1 February 2023

Unemployment data in DKI Jakarta shows an unequal class. Unbalanced data is a severe problem of modeling because it can cause prediction errors that affect the accuracy of the resulting model. Using SMOTE to handle unbalanced data will likely increase the model's accuracy. This study aims to find the best model for identifying the factors influencing the status of educated unemployment using logit and probit models and handling unbalanced data using SMOTE. The results showed that the independent variables that affect the status of educated unemployment in the logit and probit models are the same: age group and participation in training. The independent variables that affect the status of educated unemployment in the logit and probit models with SMOTE are also the same: age group, marital status, and participation in training. Unbalanced data handling using SMOTE can increase the balanced accuracy value significantly. Balanced accuracy values for the logit and probit models with SMOTE are higher than the logit and probit models without SMOTE. The logit model with SMOTE is the best because it has the highest balanced accuracy value compared to other models. According to the logit model with SMOTE, the educated unemployed in DKI Jakarta are young and have never married. There is a need for the government to play a role in improving the quality of educational institutions in producing graduates who meet company qualifications and can be hired by employers. Unemployed people who have attended the training, despite having a higher education, may also become unemployed. The training provided has not been able to reduce the unemployment rate. As a result, the government should be able to provide training to improve entrepreneurship skills while also providing capital in the form of business loans to reduce the educated unemployment.

Keywords:
SMOTE; Logit; Probit; Educated Unemployment

Format Sitasi:

F. Fatimah, A. Fitrianto, I. Indahwati, E. Erfiani, and K. N. Khikmah, "*Synthetic Minority Oversampling Technique* Pada Model Logit dan Probit Status Pengangguran Terdidik", *Jambura J. Math.*, vol. 5, No. 1, pp. 166–178, 2023, doi: https://doi.org/10.34312/jjom.v5i1.17050

1. Pendahuluan

Pendidikan pada dasarnya bertujuan untuk mempersiapkan peserta didik agar mampu menghadapi tantangan kehidupan dan mampu memasuki pasar tenaga kerja baik sebagai tenaga kerja maupun sebagai orang yang menciptakan lapangan pekerjaan (wirausaha). Jumlah penduduk yang besar seringkali dihadapkan pada masalah kurangnya lapangan pekerjaan sehingga terjadi pengangguran. Pengangguran memberikan dampak buruk, baik terhadap individu maupun masyarakat. Dampak pengangguran terhadap individu yaitu hilangnya pendapatan, kesehatan psikis/mental, dan tingkat perceraian; sedangkan dampak pengangguran terhadap masyarakat yaitu menimbulkan masalah sosial seperti meningkatnya tindak kejahatan seperti perampokan, pencurian, pencurian kendaraan bermotor, dan pembakaran [1]. Jumlah pengangguran di Indonesia berdasarkan data Sakernas 2021 tercatat sebanyak 9,2 juta jiwa atau sekitar 6,49% [2]. Pada tahun 2021, DKI Jakarta yang merupakan ibukota dan pusat perekonomian Indonesia, menempati peringkat keenam tingkat pengangguran tertinggi di Indonesia. Tingkat pengangguran terbuka di DKI Jakarta sebesar 8,50%, angka ini jauh di atas Angka Nasional yaitu 6,49%. Pengangguran di DKI Jakarta didominasi oleh pengangguran terdidik yaitu 79,18%.

Pengangguran terdidik menurut BPS adalah pengangguran dengan tingkat pendidikan SMA ke atas [3]. Pengangguran terdidik disebabkan karena kurang selarasnya perencanaan pembangunan pendidikan dengan perkembangan lapangan pekerjaan sehingga lulusan berbagai institusi pendidikan tidak terserap oleh lapangan kerja. Penyebab lain pengangguran terdidik yaitu kurang sesuainya pemilihan jenis pekerjaan yang diminati dan kurang sesuainya kualifikasi angkatan kerja yang dibutuhkan

penyedia lapangan kerja.

Pemodelan terkait pengangguran terdidik sudah pernah dilakukan diantaranya dengan menggunakan regresi berganda oleh Rosalina, *et al.* [4] yang menunjukkan bahwa variabel pendidikan dan kesempatan kerja berpengaruh terhadap pengangguran terdidik, sedangkan variabel upah tidak berpengaruh signifikan. Pemodelan lain terkait pengangguran terdidik dilakukan oleh Huda, *et al.* [5] menggunakan regresi panel dengan pendekatan *Random Effect Model*. Hasilnya menunjukkan bahwa upah minimum kabupaten, penduduk usia kerja, dan Produk Domestik Regional Bruto (PDRB) mempunyai pengaruh signifikan terhadap pengangguran terdidik di Provinsi Jawa Timur. Pemodelan lain terkait pengangguran terdidik menggunakan regresi logistik juga dapat dilihat pada [6–8]. Regresi logistik adalah model yang digunakan untuk menjelaskan hubungan antara variabel tidak bebas kategori biner dengan satu atau lebih variabel bebas. Beberapa pendekatan yang bisa digunakan untuk melakukan pendugaan parameter regresi logistik antaralain dengan pendekatan model logit dan probit. Model logit adalah model yang menggunakan fungsi logistik kumulatif, sedangkan model probit adalah model yang menggunakan fungsi normal kumulatif [9]. Pada penelitian ini dilakukan perbandingan antara model logit dan probit untuk mengetahui faktor-faktor yang berpengaruh terhadap pengangguran terdidik di DKI Jakarta. Namun persentase pengangguran terdidik dan tidak terdidik di DKI Jakarta terdapat kelas yang tidak seimbang. Data yang tidak seimbang ini menjadi masalah yang serius dalam pemodelan karena sangat berpengaruh terhadap kemampuan prediksi, akurasi, dan juga presisi [10]. Penanganan untuk data yang tidak seimbang dapat dilakukan dengan *Synthetic Minority Oversampling Technique* (SMOTE) yang merupakan pengembangan dari teknik *oversampling* untuk menyeimbangkan kelas yang tidak seimbang. SMOTE menyeimbangkan kelas yang tidak seimbang dengan cara membuat data sintetis baru dari kelas minoritas yang menyerupai data aslinya sehingga data pada kelas minoritas lebih beragam. SMOTE merupakan teknik yang paling populer dan berpengaruh dalam penanganan data yang tidak seimbang, selain itu SMOTE juga dianggap sebagai salah satu algoritma yang paling berpengaruh dalam pembelajaran mesin dan penambangan data [11].

Penanganan data tidak seimbang dengan menggunakan SMOTE sebelumnya dilakukan diantaranya oleh Ishaq, *et al.* [12], menggunakan SMOTE dalam memprediksi kelangsungan hidup pasien penyakit jantung dengan menggunakan 9 model diantaranya *decision three, adaptive boosting,* regresi logistik, *stochastic gradient, random forest, gradient boosting, extra trees, gaussian naive bayes,* dan *support verctor machine.* Data pasien penyakit jantung terdapat kelas yang tidak seimbang. Hasil penelitiannya menunjukkan bahwa *extra trees* dengan menggunakan SMOTE mampu menaikkan akurasi model secara signifikan yang semula 0,83 menjadi 0,93. Adapun pada penelitian ini dilakukan penanganan data yang tidak seimbang pada data pengangguran terdidik di DKI Jakarta dengan menggunakan SMOTE. Penelitian ini bertujuan untuk menentukan model terbaik dalam mengidentifikasi faktor-faktor yang berpengaruh terhadap status pengangguran terdidik menggunakan model logit dan probit serta melakukan penanganan data yang tidak seimbang menggunakan SMOTE.

2. Metode

2.1. Tahapan Penelitian

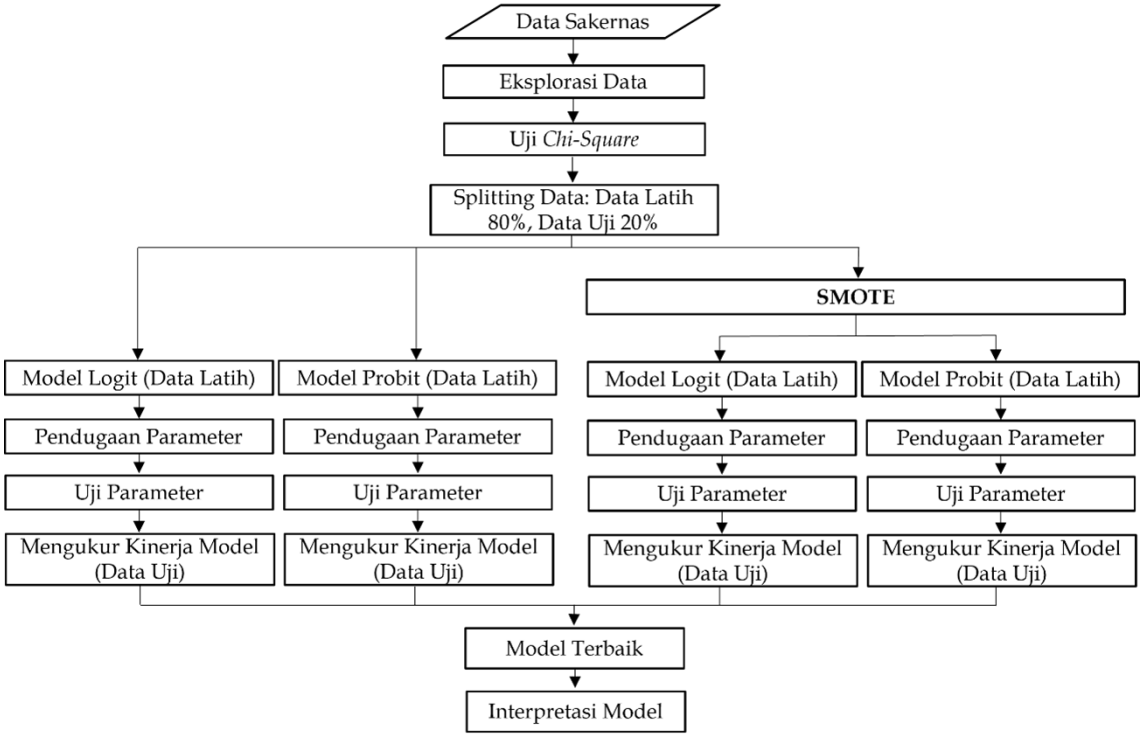
Data yang digunakan dalam penelitian ini bersumber dari hasil pendataan Sakernas Provinsi DKI Jakarta Agustus 2021. Penelitian ini menggunakan dua metode analisis yaitu analisis deskriptif dan analisis inferensia. Analisis deskriptif digunakan untuk menggambarkan karakteristik responden yang menjadi objek penelitian, sedangkan analisis inferensia digunakan untuk membangun model dan mengetahui faktor apa saja yang berpengaruh dalam menentukan status seseorang masuk ke dalam pengangguran terdidik atau tidak. Analisis inferensia menggunakan analisis logit dan probit serta menggunakan SMOTE untuk penanganan data yang tidak seimbang.

Sebelum dilakukan pemodelan, dilakukan uji asosiasi untuk mengetahui ada atau tidaknya hubungan masing-masing variabel bebas terhadap variabel tidak bebasnya. Pengujian dilakukan dengan menggunakan Uji *Chi-Square* [13]. Langkah selanjutnya dilakukan *splitting data*. *Splitting data* adalah membagi data menjadi dua bagian yaitu data latih dan data uji. Data latih adalah data yang digunakan untuk melatih model, sedangkan data uji adalah data yang digunakan untuk mengevaluasi model. Rasio *splitting data* menggunakan Prinsip Pareto yang telah digunakan secara luas yaitu membagi data menjadi dua bagian: data latih sebanyak 80% dari keseluruhan data dan data uji sebanyak 20% dari keseluruhan data [14]. Langkah selanjutnya dilakukan pemodelan dengan menggunakan model logit, probit dan penanganan data yang tidak seimbang menggunakan SMOTE. Setelah dilakukan pemodelan, dilakukan uji parameter yang meliputi uji kesuaian model, uji simultan, dan uji parsial. Pemilihan model terbaik adalah model yang memiliki nilai *balanced accuracy* tertinggi.

Variabel-variabel yang digunakan dalam penelitian ini disajikan pada Tabel 1 dan tahapan penelitian disajikan pada Gambar 1.

Tabel 1. Variabel-variabel yang digunakan

Variabel	Kategori	*Dummy*	Referensi
Pengangguran	Terdidik	1	
	Tidak terdidik	0	
Jenis kelamin	Laki-laki	1	[15]
	Perempuan	0	
Kelompok umur	15-25 tahun	1	[7, 8, 15, 16]
	26-45 tahun	10	
	$>$ 45 tahun	0	
Status perkawinan	Pernah kawin	1	[6–8, 15, 16]
	Belum pernah kawin	0	
Status kepala rumah tangga (KRT)	KRT	1	[6, 16]
	Bukan KRT	0	
Pengalaman kerja	Pernah bekerja	1	[6, 7, 16]
	Belum pernah bekerja	0	
Pelatihan	Pernah pelatihan	1	[6, 7]
	Belum pernah pelatihan	0	



Gambar 1. *Flowchart* Tahapan penelitian

Selanjutnya dipaparkan beberapa konsep penting berkaitan dengan penelitian ini. Konsep-konsep penting yang dipaparkan berikut diperlukan pada pembahasan hasil penelitian.

2.2. Model Logit dan Probit

Model regresi logistik adalah model yang digunakan untuk menjelaskan hubungan antara variabel tidak bebas kategori biner dengan satu atau lebih variabel bebas [9]. Persamaan umum dari regresi logistik adalah

$$\pi\left(x\right) = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}} \tag{1}$$

dengan $i = 1,\ 2,\ \ldots,\ p$.

Pendugaan parameter regresi logistik salah satunya adalah dengan model logit. Model logit menggunakan fungsi kumulatif logistik. Agar model menjadi linier dilakukan transformasi logit terhadap $\pi\left(x\right)$. Persamaan dari model logit adalah

$$g\left(x\right) = ln\left[\frac{\pi\left(x\right)}{1 - \pi\left(x\right)}\right] = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \tag{2}$$

dengan $i = 1,\ 2,\ \ldots,\ p$. Interpretasi pada model logit akan lebih bermakna jika menggunakan *odds ratio* sebagai ukuran asosiasi. *Odds ratio* merupakan perbandingan peluang suatu kejadian terjadi dengan peluang kejadian tidak terjadi. Rumus dari *odds ratio* adalah

$$OR = \frac{odds_1}{odds_0} = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} \tag{3}$$

Pendekatan lain dalam menduga parameter regresi logistik adalah dengan model probit. Model probit diperkenalkan pertama kali oleh Bliss pada tahun 1935. Model probit menggunakan fungsi kumulatif normal. Agar model menjadi linier dilakukan transformasi dengan *probability unit.* Persamaan model probit setelah dilakukan transformasi adalah

$$\text{Œ}^{-1}P\left(y = 1 \mid x\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \tag{4}$$

dengan $i = 1,\ 2,\ \ldots,\ p$.

2.3. Uji Parameter

Dilakukan tiga pengujian yaitu uji kesuaian model, uji simultan, dan uji parsial. Uji kesuaian model digunakan untuk menguji apakah model yang terbentuk sudah tepat atau tidak. Model dapat dikatakan sudah tepat jika tidak ada perbedaan signifikan antara hasil prediksi model dengan nilai observasinya.

Uji simultan atau *likelihood ratio test* digunakan untuk mengetahui peranan seluruh variabel bebas secara simultan terhadap variabel tidak bebasnya. Dalam hal ini, digunakan hipotesis berikut, yaitu

$H_0 :\ \beta_0 = \beta_1 = \cdots = \beta_p = 0$ (tidak ada pengaruh variabel bebas secara simultan terhadap variabel tidak bebas)

$H_1 :$ minimal satu $\beta_j \neq 0$, dengan $j = 0,\ 1,\ \ldots,\ p$ (ada pengaruh paling sedikit satu variabel bebas terhadap variabel tidak bebas).

Statistik uji yang digunakan yaitu

$$G = -2ln\left[\frac{likelihood\ tanpa\ variabel\ penjelas}{likelihood\ dengan\ variabel\ penjelas}\right]. \tag{5}$$

Keputusan H_0 akan ditolak jika $G > \lambda_{(\alpha,p)}$ atau jika $p - value < \alpha$.

Uji parsial atau uji Wald digunakan untuk mengetahui variabel-variabel bebas mana saja yang berpengaruh terhadap variabel tidak bebasnya.

Hipotesis yang digunakan:

$H_0 :\ \beta_j = 0$, dengan $j = 0,\ 1,\ \ldots,\ p$

$H_1 :\ \beta_j \neq 0$, dengan $j = 0,\ 1,\ \ldots,\ p$

dengan statistik uji, yaitu

$$W^2 = \left[\frac{\widehat{\beta}_j}{\widehat{SE}\widehat{\beta}_j}\right]^2 \tag{6}$$

Keputusan H_0 akan ditolak jika $W^2 > \lambda_{(1)}$ atau jika $p - value < \alpha$.

2.4. Data Tidak Seimbang

Data tidak seimbang adalah data dengan jumlah pengamatan pada satu kelas secara signifikan lebih kecil dibanding kelas yang lain [17]. Suatu data disebut tidak seimbang jika rasio ketidakseimbangan 1:4 hingga 1:100 [18]. Data yang tidak seimbang perlu ditangani karena data tidak seimbang dapat mengakibatkan kesalahan prediksi sehingga dapat mempengaruhi akurasi model [19]. Teknik yang dapat digunakan untuk mengatasi masalah data tidak seimbang yaitu: teknik *undersampling* dan teknik *oversampling* [20]. SMOTE adalah sebuah teknik yang kurang lebih sama dengan teknik *oversampling*. SMOTE tidak hanya menduplikasi data yang sama akan tetapi SMOTE akan membuat data baru yang menyerupai data asli dari kelas minoritas untuk menyeimbangkan data, sehingga data baru dari kelas minoritas jauh lebih beragam [18, 20]. Pengunaan SMOTE dalam data berdimensi rendah (jumlah variabel jauh lebih sedikit dibanding jumlah amatan) sangat efektif dalam penanganan kelas yang tidak seimbang, sedang pada data berdimensi tinggi (jumlah variabel jauh lebih banyak dibanding jumlah amatan) penggunaan SMOTE kurang efektif [18].

2.5. Pengukuran Kinerja Model

Ada beberapa metode untuk mengukur kinerja model, misalnya dengan akurasi, sensitivitas, spesifisitas, dan skor *Area Under Curve* (AUC) [17]. Penggunaan akurasi, sensitivitas, spesifisitas maupun skor AUC dalam mengukur kinerja model pada data yang tidak seimbang kurang tepat [18]. Ada beberapa metode untuk mengukur kinerja model pada data tidak seimbang diantaranya dengan *balanced accuracy* [21]. Rumus *balanced accuracy* yang digunakan yaitu

$$\text{Balanced Accuracy} = \frac{\text{Sensitivitas} + \text{Spesifisitas}}{2} \tag{7}$$

dengan $sensitivitas = \frac{TP}{TP+FN}$ dan $spesifisitas = \frac{TN}{TN+FP}$. Perhitungan *balanced accuracy* dengan menggunakan nilai yang ada pada Tabel 2.

Tabel 2. *Confusion matrix*

Prediksi	Kenyataan	
	Pengangguran terdidik	Pengangguran tidak terdidik
Pengangguran terdidik	*True Positive* (TP)	*False Positive* (FP)
Pengangguran tidak terdidik	*False Negative* (FN)	*True Negative* (TN)

3. Hasil dan Pembahasan

3.1. Analisis Deskriptif

Berikut adalah gambaran umum responden dalam penelitian ini, berdasarkan dari data sampel 635 pengangguran hasil pendataan Sakernas Agustus 2021 DKI Jakarta, persentase pengangguran terdidik yaitu sebesar 79,18% sedangkan persentase pengangguran tidak terdidik yaitu sebesar 20,82%. Pengangguran dilihat menurut jenis kelaminnya, persentase pengangguran laki-laki jauh lebih tinggi dibanding pengangguran perempuan, sedangkan jika dilihat dari kelompok umurnya persentase pengangguran didominasi oleh kelompok umur 15-25 tahun dan 26-45 tahun.

Pengangguran dilihat dari status perkawinannya, persentase pengangguran yang belum pernah kawin jauh lebih tinggi dibanding yang pernah kawin, sedang jika dilihat dari status dalam rumah pengangguran didominasi oleh mereka yang berstatus bukan KRT. Pengangguran jika dilihat dari pengalaman kerjanya, maka persentase pengangguran yang belum pernah bekerja sedikit lebih tinggi dibanding yang pernah bekerja. Pengangguran jika dilihat dari pernah atau tidaknya mengikuti pelatihan, maka persentase pengangguran yang tidak pernah mengikuti pelatihan jauh lebih tinggi dibanding yang pernah mengikuti pelatihan. Selengkapnya bisa dilihat pada Tabel 3.

Tabel 3. Karakterisitik pengangguran di DKI Jakarta

Variabel	Kategori	Persentase
Jenis Kelamin	Laki-laki	63,45%
	Perempuan	36,55%
Kelompok Umur	15-25 tahun	46,72%
	26-45 tahun	42,20%
	$>$ 45 tahun	11,08%
Status Kawin	Pernah Kawin	31,98%
	Belum Pernah Kawin	68,02%
Status KRT	Kepala Rumah Tangga	13,85%
	Bukan Kepala Rumah Tangga	86,15%
Pengalaman Kerja	Pernah Kerja	47,02%
	Belum Pernah Kerja	52,98%
Pelatihan	Pernah Pelatihan	14,86%
	Tidak Pernah Pelatihan	85,14%

Karakteristik pengangguran jika dilihat menurut status penganggurannya, disajikan pada Tabel 4.

Tabel 4. Karakterisitik pengangguran menurut status pengangguran di DKI Jakarta

Variabel	Kategori	Status pengangguran		Total
		Terdidik	Tidak terdidik	
Jenis Kelamin	Laki-laki	79,50%	20,50%	100%
	Perempuan	78,70%	21,30%	100%
Kelompok Umur	15-25 tahun	89,20%	10,80%	100%
	26-45 tahun	72,00%	28,00%	100%
	$>$ 45 tahun	64,00%	36,00%	100%
Status Kawin	Pernah Kawin	68,40%	31,60%	100%
	Belum Pernah Kawin	84,20%	15,80%	100%
Status KRT	Kepala Rumah Tangga	68,10%	31,90%	100%
	Bukan Kepala Rumah Tangga	81,00%	19,00%	100%
Pengalaman Kerja	Pernah Kerja	80,90%	19,10%	100%
	Belum Pernah Kerja	77,70%	22,30%	100%
Pelatihan	Pernah Pelatihan	98,90%	1,10%	100%
	Tidak Pernah Pelatihan	75,70%	24,30%	100%

Dilihat menurut jenis kelaminnya persentase pengangguran terdidik laki-laki sedikit lebih tinggi dibanding perempuan, sedangkan jika dilihat dari kelompok umurnya persentase pengangguran terdidik kelompok umur 15-25 tahun paling tinggi dibanding kelompok umur lainnya. Pengangguran terdidik dilihat dari status perkawinannya,

persentase pengangguran terdidik yang belum pernah kawin lebih tinggi dibanding yang pernah kawin, sedang jika dilihat dari status dalam rumah pengangguran terdidik yang berstatus bukan KRT lebih tinggi dibanding yang merupakan KRT. Pengangguran terdidik jika dilihat dari pengalaman kerjanya, maka persentase pengangguran terdidik yang pernah bekerja sedikit lebih tinggi dibanding yang belum pernah bekerja. Pengangguran terdidik jika dilihat dari pernah atau tidaknya mengikuti pelatihan, maka persentase pengangguran terdidik yang pernah mengikuti pelatihan lebih tinggi dibanding yang belum pernah mengikuti pelatihan.

3.2. Hasil Uji Asosiasi (Uji Chi-square)

Sebelum dilakukan pemodelan, dilakukan uji asosiasi untuk mengetahui ada atau tidaknya hubungan masing-masing variabel bebas terhadap variabel tidak bebasnya. Variabel bebas memiliki hubungan dengan variabel tidak bebas jika *p-value* kurang dari $0,05$. Hasil pengujian selengkapnya disajikan pada Tabel 5.

Tabel 5. *Chi-square value*, derajat bebas, *p-value* hasil uji asosiasi antara variabel bebas dengan variabel tidak bebas (status pengangguran terdidik)

Variabel bebas	*Chi-square value*	Derajat bebas	*P-value*	Keterangan
Jenis kelamin	0,0004	1	0,9850	Tidak signifikan
Kelompok umur	43,7450	2	3,17 e-10*	Signifikan
Status kawin	29,9520	1	4,42 e-8*	Signifikan
Status KRT	7,3807	1	0,0066*	Signifikan
Pengalaman kerja	2,47 e-30	1	1	Tidak signifikan
Pelatihan	20,1020	1	7,34 e-6*	Signifikan

Hasil uji asosiasi dengan menggunakan uji *Chi-square* menunjukkan bahwa variabel kelompok umur, status kawin, status KRT, dan pelatihan memiliki asosiasi dengan status pengangguran terdidik di DKI Jakarta, sedangkan variabel jenis kelamin dan pengalaman kerja tidak memiliki asosiasi dengan status pengangguran terdidik di DKI Jakarta. Variabel bebas yang dimasukkan ke dalam model antaralain: kelompok umur, status kawin, status KRT, dan keikutsertaan dalam pelatihan.

3.3. Splitting Data

Data dibagi menjadi dua bagian yaitu 80% (508 data pengangguran) sebagai data latih dan 20% (127 data pengangguran) sebagai data uji. Dari 508 data pengangguran yang menjadi data latih, 396 pengangguran dikategorikan sebagai pengangguran terdidik dan sebanyak 112 pengangguran dikategorikan sebagai pengangguran tidak terdidik.

3.4. Pemodelan

Pemodelan dilakukan dengan menggunakan model logit, probit, logit dengan SMOTE, dan probit dengan SMOTE. Hasil pemodelan logit dan probit disajikan pada Tabel 6, sedangkan hasil pemodelan logit dan probit dengan SMOTE disajikan Tabel 7.

Variabel bebas yang berpengaruh terhadap status pengangguran terdidik pada model logit dan probit sama yaitu kelompok umur (umur 15-25 tahun vs umur $>$ 45 tahun) dan keikutsertaan dalam pelatihan (pernah pelatihan vs tidak pernah pelatihan), sedang pada model logit dan probit dengan SMOTE, variabel bebas yang berpengaruh terhadap status pengangguran terdidik juga sama yaitu kelompok umur (umur 15-25

Tabel 6. Hasil pemodelan logit dan probit

	Model logit		Model probit	
	Koefisien penduga parameter	*P-value*	Koefisien penduga parameter	*P-value*
Intercept	0,3613	0,3704	0,2431	0,3209
Umur 15-25 tahun vs umur > 45 tahun	1,5710	0,0004*	0,8843	0,0008*
Umur 26-45 tahun vs e umur > 45 tahun	0,4687	0,1765	0,2795	0,1859
Pernah kawin vs belum pernah kawin	-0,4189	0,1766	-0,2565	0,1675
KRT vs bukan KRT	0,3230	0,3636	0,2001	0,3489
Pernah pelatihan vs tidak pernah pelatihan	2,2023	0,0003*	1,1011	6,03 e-5*

Tabel 7. Hasil pemodelan logit dan probit dengan SMOTE

	Model logit dengan SMOTE		Model probit dengan SMOTE	
	Koefisien penduga parameter	*P-value*	Koefisien penduga parameter	*P-value*
Intercept	-0,4501	0,1775	-0,2634	0,1882
Umur 15-25 tahun vs umur > 45 tahun	1,1815	0,0009*	0,7023	0,0009*
Umur 26-45 tahun vs e umur > 45 tahun	0,3523	0,2571	0,2074	0,2586
Pernah kawin vs belum pernah kawin	-0,9760	7,8 e-5*	-0,5906	7,9 e-5*
KRT vs bukan KRT	0,3677	0,2285	0,2249	0,2152
Pernah pelatihan vs tidak pernah pelatihan	2,5352	3,6 e-7*	1,4030	4,6 e-8*

tahun vs umur > 45 tahun), status perkawinan (pernah kawin vs belum pernah kawin), dan keikutsertaan dalam pelatihan (pernah pelatihan vs tidak pernah pelatihan).

3.5. Pemilihan Model Terbaik

Pemilihan model terbaik dalam penelitian ini dengan melihat nilai *balanced accuracy* tertinggi. Perbandingan *balanced accuracy* untuk keempat model disajikan pada Tabel 8. Model logit dan probit menghasilkan nilai *balanced accuracy* yang sama, sedangkan model logit dengan SMOTE menghasilkan *balanced accuracy* yang sedikit lebih tinggi dibandingkan probit dengan SMOTE. Model terbaik dalam penelitian ini yaitu model logit dengan SMOTE karena memiliki *balanced accuracy* tertinggi dibanding ketiga model lainnya.

Tabel 8. *Balanced accuracy* model logit, probit, logit dengan SMOTE, dan probit dengan SMOTE

Model	*Balanced accuracy*
Logit	0,5200
Probit	0,5200
Logit dengan SMOTE	0,7344
Probit dengan SMOTE	0,7266

Persamaan model logit dengan SMOTE yang terbentuk adalah

$$SPT = -0,4501 - 0,9760 * K + 0,3677 * KRT + 1,1815 * U15 - U25thn \quad (8)$$
$$+0,3523 * U26 - U45thn + 2,5352 * PP.$$

dengan

SPT	=	Status Pengangguran terdidik
KRT	=	Kepala Rumah Tangga
K	=	Kawin
U	=	Umur
PP	=	Pernah Pelatihan

Hasil pengujian dengan Uji Hosmer Lemeshow menghasilkan $\hat{C} = 6,7751$ dengan $p - value = 0,2379$, karena $p - value > 0,05$ maka dapat disimpulkan model yang terbentuk sudah tepat/tidak terdapat perbedaan yang signifikan antara data observasi dengan hasil prediksi modelnya. Langkah selanjutnya yaitu dilakukan pengujian secara simultan maupun parsial terhadap model. Hasil pengujian secara simultan diperoleh statistik uji $G = 109,2068$, karena statistik uji $G > \lambda_{(0,05,4)}$ maka dapat disimpulkan bahwa variabel-variabel bebas berpengaruh secara simultan terhadap status pengangguran terdidik. Hasil pengujian secara parsial menunjukkan bahwa variabel bebas yang berpengaruh terhadap status pengangguran terdidik diantaranya: kelompok umur (umur 15-25 tahun vs umur $>$ 45 tahun), status perkawinan (pernah kawin vs belum pernah kawin), dan keikutsertaan dalam pelatihan (pernah pelatihan vs tidak pernah pelatihan). Selengkapnya disajikan pada Tabel 9.

Tabel 9. Nilai estimasi koefisien penduga parameter dan nilai o*dds ratio* pada model logit dengan SMOTE

	Estimasi koefisien	*Odds ratio*
Intercept	-0,4501	0,6376
Umur 15-25 tahun vs umur $>$ 45 tahun	1,1815	3,2593
Pernah kawin vs belum pernah kawin	-0,9760	0,3768
Pernah pelatihan vs tidak pernah pelatihan	2,5352	12,6190

Berdasarkan Tabel 9 diketahui bahwa *odds ratio* variabel *dummy* kelompok umur muda (15-25 tahun) yaitu 3,2593. Hal ini berarti seorang pengangguran yang berumur 15-25 tahun memiliki risiko untuk menjadi pengangguran terdidik 3,2593 kali dibanding seorang pengangguran yang berumur lebih dari 45 tahun. Hal ini sejalan dengan penelitian yang dilakukan oleh Aulia [7, 8] yang menyatakan bahwa pengangguran terdidik cenderung terjadi oleh seorang pengangguran yang berusia muda dibanding seorang pengangguran yang berusia lebih tua.

Odds ratio variabel *dummy* status kawin yaitu 0,3768. Hal ini berarti seorang pengangguran yang berstatus pernah kawin memiliki risiko untuk menjadi pengangguran terdidik 0,3768 kali dibandingkan seorang pengangguran yang berstatus belum pernah kawin atau dengan kata lain seorang pengangguran yang berstatus belum pernah kawin memiliki risiko untuk menjadi pengangguran terdidik 1/0.3768 = 2,6539 kali dibandingkan seorang pengangguran yang berstatus pernah kawin. Hal ini sejalan dengan hasil penelitian yang dilakukan oleh Aulia [7, 8], dan Alharis [6] yang menyatakan bahwa seorang pengangguran yang belum pernah kawin cenderung untuk

menjadi pengangguran terdidik dibandingkan seorang pengangguran yang sudah pernah kawin.

Odds ratio variabel *dummy* pelatihan yaitu 12,6190. Hal ini berarti seorang pengangguran yang pernah mengikuti pelatihan memiliki risiko 12,6190 kali untuk menjadi pengangguran terdidik dibandingkan seorang pengangguran yang belum pernah mengikuti pelatihan. Hal ini juga sejalan dengan penelitian Aulia [7] dan Alharis [6] yang menyatakan bahwa seorang pengangguran yang mengikuti pelatihan justru yang cenderung menjadi pengangguran terdidik dibanding seorang pengangguran yang tidak pernah mengikuti pelatihan. Persentase pengangguran terdidik yang pernah pelatihan sekitar 18,56%. Karakteristik pengangguran terdidik yang pernah mengikuti pelatihan yaitu: mayoritas berusia 15-45 tahun; 59,83% merupakan laki-laki; 66,72% berstatus belum kawin; dan 80,05% bukan merupakan KRT. Penyebab fenomena ini kemungkinan disebabkan karena kurang efektifnya Balai Latihan Kerja dalam menekan angka pengangguran [22, 23]. Pengangguran yang pernah mendapat pelatihan kerja meskipun berpendidikan tinggi tidak sepenuhnya mampu terserap dalam perusahaan karena perbedaan keahlian yang dimiliki dengan kebutuhan perusahaan, atau karena belum mampu menciptakan usaha mandiri dikarenakan keterbatasan modal yang dimiliki.

4. Kesimpulan

Hasil penelitian ini menunjukkan variabel bebas yang berpengaruh terhadap status pengangguran terdidik pada model logit dan probit sama yaitu kelompok umur (umur 15-25 tahun vs umur $>$ 45 tahun) dan keikutsertaan dalam pelatihan (pernah pelatihan vs tidak pernah pelatihan), sedang pada model logit dan probit dengan SMOTE, variabel bebas yang berpengaruh terhadap status pengangguran terdidik juga sama yaitu kelompok umur (umur 15-25 tahun vs umur $>$ 45 tahun), status perkawinan (pernah kawin vs belum pernah kawin), dan keikutsertaan dalam pelatihan (pernah pelatihan vs tidak pernah pelatihan). Penanganan data tidak seimbang dengan menggunakan SMOTE mampu menaikkan nilai *balanced accuracy* secara signifikan. Hal ini terlihat dari nilai *balanced accuracy* untuk model baik logit dan probit dengan SMOTE lebih tinggi dibanding model logit dan probit tanpa SMOTE. Model logit dengan SMOTE merupakan model terbaik karena memiliki nilai *balanced accuracy* tertinggi dibanding model lain. Berdasarkan model logit dengan SMOTE, pemerintah perlu mewaspadai adanya fenomena pengangguran terdidik yang ternyata berasal dari kelompok umur muda dan berstatus belum pernah kawin, hal ini menunjukkan kurangnya kesempatan kerja dalam menyerap lulusan institusi pendidikan SMA ke atas sehingga perlu adanya peran pemerintah dalam meningkatkan kualitas institusi pendidikan dalam menghasilkan lulusan yang sesuai kualifikasi perusahaan sehingga mampu terserap oleh perusahaan penyedia lapangan kerja. Berdasarkan model tersebut juga, pengangguran yang pernah mengikuti pelatihan walaupun berpendidikan tinggi ternyata juga berpotensi untuk menjadi pengangguran. Bekal pelatihan yang pernah diikuti ternyata belum mampu menekan angka pengangguran, olehkarenanya pemerintah hendaknya mampu menyediakan pelatihan-pelatihan yang dapat meningkatkan kemampuan berwirausaha dan sekaligus menyediakan modal berupa kredit usaha sehingga dapat menekan angka pengangguran terdidik.

Referensi

[1] M. Kassem, A. Ali, and M. Audi, "Unemployment Rate, Population Density and Crime Rate in Punjab (Pakistan): An Empirical Analysis," *Bulletin of Business and Economics*, vol. 8, no. 2, pp. 92–104, 2019.

[2] BPS (Badan Pusat Statistik), *Keadaan Angkatan Kerja di Indonesia Agustus 2021*. Jakarta: Badan Pusat Statistik, 2021.

[3] ——, *Ringkasan Eksekutif Informasi Ketenagakerjaan Provinsi Sumatera Barat 2015*. Padang: Badan Pusat Statistik Provinsi Sumatera Barat, 2016.

[4] R. Rosalina, P. H. Prihanto, and E. Achmad, "Faktor-faktor yang mempengaruhi tingkat pengangguran terdidik di Provinsi Jambi," *e-Jurnal Ekonomi Sumberdaya dan Lingkungan*, vol. 6, no. 3, pp. 123–133, 2017.

[5] M. M. Huda, I. W. Subagiarta, and M. Adenan, "Determinan Pengangguran Terdidik Jawa Timur," *e-Journal Ekonomi Bisnis dan Akuntansi*, vol. 5, no. 1, pp. 48–52, may 2018, doi: 10.19184/ejeba.v5i1.7733.

[6] F. A. Alharis and A. F. Yuniasih, "Determinan Pengangguran Usia Muda Terdidik di Provinsi Banten Tahun 2020," in *Seminar Nasional Official Statistics*, vol. 2022, no. 1, nov 2022, pp. 53–62, doi: 10.34123/semnasoffstat.v2022i1.1153.

[7] M. F. Aulia, "Determinan Pengangguran Terdidik di Jawa Timur," *Jurnal Ilmiah Mahasiswa FEB Universitas Brawijaya*, vol. 5, no. 2, 2017.

[8] N. R. Aulia and L. Yuliana, "Determinan Pengangguran Terdidik di Wilayah Perkotaan Perdesaan dan Wilayah Perkotaan Provinsi Kepulauan Riau Tahun 2021," *Seminar Nasional Official Statistics*, vol. 2022, no. 1, pp. 275–284, nov 2022, doi: 10.34123/semnasoffstat.v2022i1.1367.

[9] A. Agresti, *An Introduction to Categorical Data Analysis*, 3rd ed. New Jersey: John Wiley & Sons, Inc., 2018.

[10] C. Salas-Eljatib, A. Fuentes-Ramirez, T. G. Gregoire, A. Altamirano, and V. Yaitul, "A study on the effects of unbalanced data when fitting logistic regression models in ecology," *Ecological Indicators*, vol. 85, pp. 502–508, feb 2018, doi: 10.1016/j.ecolind.2017.10.030.

[11] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Systems*, vol. 98, pp. 1–29, apr 2016, doi: 10.1016/j.knosys.2015.12.006.

[12] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39 707–39 716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[13] D. Forsyth, *Probability and Statistics for Computer Science*. Cham: Springer International Publishing, 2018, doi: 10.1007/978-3-319-64410-3.

[14] V. R. Joseph, "Optimal ratio for data splitting," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, pp. 531–538, aug 2022, doi: 10.1002/sam.11583.

[15] F. F. Adyaksa, "Analisis Faktor-Faktor yang Mempengaruhi Pengangguran Terdidik di Indonesia Tahun 2018," *E-Jurnal Ilmu Ekonomi dan Bisnis Universitas Brawijaya*, vol. 8, no. 2, pp. 1–10, 2019.

[16] M. V. Makung, R. Hadi, Y. Rosaripatria, and S. I. Oktora, "Determinan Pengangguran Terdidik Di Provinsi Nusa Tenggara Timur (NTT) Tahun 2018 Menggunakan Regresi Logistik Bine," *Jurnal Statistika Universitas Muhammadiyah Semarang*, vol. 9, no. 2, pp. 64–78, dec 2021, doi: 10.26714/jsunimus.9.2.2021.64-78.

[17] T. Beysolow II, *Introduction to Deep Learning Using R*. Berkeley, CA: Apress, 2017, doi: 10.1007/978-1-4842-2734-3.

[18] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, nov 2016, doi: 10.1007/s13748-016-0094-0.

[19] X. Chao, G. Kou, Y. Peng, and A. Fernández, "An efficiency curve for evaluating imbalanced classifiers considering intrinsic data characteristics: Experimental analysis," *Information Sciences*, vol. 608, pp. 1131–1156, aug 2022, doi: 10.1016/j.ins.2022.06.045.

[20] Z. Zhang, H. Liu, D. Chen, J. Zhang, H. Li, M. Shen, Y. Pu, Z. Zhang, J. Zhao, and J. Hu, "SMOTE-based method for balanced spectral nondestructive detection of moldy apple core," *Food Control*, vol. 141, p. 109100, nov 2022, doi: 10.1016/j.foodcont.2022.109100.
[21] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, jul 2019, doi: 10.1016/j.patcog.2019.02.023.
[22] N. Ismi, *Efektifitas Balai Latihan Kerja dalam Mengurangi Pengangguran di Kabupaten Bone*. Skripsi: Universitas Muhammadiyah Makassar, 2020.
[23] M. N. Pratama, N. Widowati, and M. Maesaroh, "Efektivitas Program Pelatihan Kerja UPTD Balai Latihan Kerja Dinas Tenaga Kerja Kota Semarang," *Journal of Public Policy and Management Review*, vol. 10, no. 2, pp. 104–116, 2021, doi: 10.14710/jppmr.v10i2.30593.