IJSA (2021), Vol. 5, No. 1, 75-91
DOI: https://doi.org/10.29244/ijsa.v5i1p75-91

Indonesian Journal of Statistics and Its Applications
Available at https://ijsa.stats.id
e-ISSN: 2599-0802

Exploration of Obesity Status of Indonesia Basic Health Research 2013 With Synthetic Minority Over-Sampling Techniques*

Eksplorasi Status Obesitas Riset Kesehatan Dasar 2013 Indonesia dengan Teknik *Synthetic Minority Over-Sampling*

Sri Astuti Thamrin[1‡], Dian Sidik[2], Hedi Kuswanto[1], Armin Lawi[3], and Ansariadi[2]

[1]Departemen Statistika, Universitas Hasanuddin, Indonesia
[2]Departemen Epidemiologi, Universitas Hasanuddin, Indonesia
[3]Departemen Matematika, Universitas Hasanuddin, Indonesia
[‡]corresponding author: tuti@unhas.ac.id

Abstract

The accuracy of the data class is very important in classification with a machine learning approach. The more accurate the existing data sets and classes, the better the output generated by machine learning. In fact, classification can experience imbalance class data in which each class does not have the same portion of the data set it has. The existence of data imbalance will affect the classification accuracy. One of the easiest ways to correct imbalanced data classes is to balance it. This study aims to explore the problem of data class imbalance in the medium case dataset and to address the imbalance of data classes as well. The Synthetic Minority Over-Sampling Technique (SMOTE) method is used to overcome the problem of class imbalance in obesity status in Indonesia 2013 Basic Health Research (RISKESDAS). The results show that the number of obese class (13.9%) and non-obese class (84.6%). This means that there is an imbalance in the data class with moderate criteria. Moreover, SMOTE with over-sampling 600% can improve the level of minor classes (obesity). As consequence, the classes of obesity status balanced. Therefore, SMOTE technique was better compared to without SMOTE in exploring the obesity status of Indonesia RISKESDAS 2013.

Keywords: Imbalanced data, machine learning, obesity status, SMOTE.

* Received: Nov 2020, Reviewed: Mar 2021, Published: Mar 2021

1 Pendahuluan

Obesitas dapat terjadi karena adanya ketidakseimbangan jumlah makanan yang masuk dibandingkan dengan pengeluaran energi yang dilakukan oleh tubuh ([RISKESDAS], 2013). Di Asia Tenggara, jumlah kasus penderita kasus gizi buruk dan gizi kurang masih cukup tinggi tetapi jumlah kasus kegemukan dan obesitas semakin meningkat (ASEAN / UNICEF / WHO Regional Report, 2016). Di Indonesia, angka prevalensi obesitas terus mengalami peningkatan khususnya pada wanita dewasa. Dari tahun 2007 dan 2010, angka prevalensi obesitas pada wanita dewasa masing-masing sebesar 14,8%, 15,5% dan meningkat tajam sebesar 32,9% pada tahun 2013. Sementara, laki-laki dewasa mengalami obesitas pada tahun 2007 sebanyak 13,9%, kemudian sedikit menurun ke angka 7,8% pada tahun 2010 dan meningkat lg pada tahun 2013 sebesar 19,7% ([KEMENKES RI], 2014). Oleh karena itu penting untuk di eksplorasi data status obesitas di Indonesia.

Status obesitas dapat dieksplorasi dengan menggunakan teknik pembelajaran mesin (*machine learning*). Dalam klasifikasi pada pembelajaran mesin, akurasi dari kelas data sangat penting (Ente et al., 2020; Selya & Anshutz, 2018; Witten et al., 2011). Indikator luaran yang baik dari pembelajaran mesin dapat dilihat dari akurasi dataset dan kelas yang ada. Dataset akurat dicapai jika kelas data tidak mengalami ketidakseimbangan.

Masalah ketidakseimbangan data (Branco et al., 2016; Krawczyk, 2016, Sun et al., 2009) terjadi dalam tugas klasifikasi setiap kali jumlah pengamatan milik salah satu kelas, kelas mayoritas, melebihi jumlah pengamatan milik salah satu kelas lain, kelas minoritas. Algoritma klasifikasi tradisional rentan terhadap keberadaan data yang tidak seimbang, dan cenderung menampilkan bias terhadap kelas mayoritas dengan mengorbankan kemampuan diskriminasi kelas minoritas. Efek negatif pada kinerja klasifikasi ini diperburuk oleh adanya faktor kesulitan dataset tambahan, seperti disjungsi kecil (Jo & Japkowicz, 2004) atau jumlah observasi data latih yang tidak mencukupi (Chen & Wasikowski, 2008), yang dapat menyebabkan model *over-fitting*.

Data tidak seimbang dapat diatasi dengan tiga pendekatan (Yap et al., 2014), yaitu pendekatan pada level data, *classifier* dan *ensemble*. Teknik *over-sampling* dan *under-sampling* digunakan pada pendekatan level data. Selain itu, pada pendekatan level data ini, kecondongan distribusi kelas data dapat diperbaiki melalui data buatan (*synthetic*). Pada level data ini dapat digunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*) (N. V. Chawla et al., 2002). Pendekatan level *classifier* berfokus pada pengklasifikasi (*classifier*) kelas minoritas menggunakan operasi algoritma (Zhang et al., 2011). Kemudian, pendekatan *ensemble* dilakukan perbaikan algoritma pengklasifikasi.

SMOTE merupakan algoritma dengan pendekatan *over-sampling* yaitu data buatan untuk kelas data minoritas dibangkitkan agar terjadi keseimbangan proporsi kelas data mayor dan minor (Alghamdi et al., 2017; N. V. Chawla et al., 2002; Mustaqim et al., 2019). Keuntungan dari pendekatan ini adalah mampu mengatasi masalah ketidakseimbangan data kelas karena sejumlah data buatan ditambahkan pada kelas minoritas yang membuat distribusi data dalam kelas menjadi seimbang. Ada tiga kasus ketidakseimbangan kelas data yaitu ringan (20-40%), sedang (1-20%), dan ekstrem (<1%), bergantung pada proporsi kelas minoritas ke seluruh kumpulan data. Oleh karena itu dalam studi ini akan dieksplorasi masalah ketidakseimbangan data kelas untuk kasus dataset sedang dan mengatasi ketidakseimbangan data kelas dari status obesitas Survei Riset Kesehatan Dasar (RISKESDAS) 2013 Indonesia.

tersebut, k-tetangga terdekat (*k-nearest neighbor*) digunakan (N V Chawla et al , 2002) Pada studi ini, teknik SMOTE dilakukan dengan cara menghitung jarak antar sampel kelas minornya Jarak antar sampel minor ini dilakukan dengan metode *Modify Value Difference Metric* (MVDM) (Cost & Salzberg, 1993) Adapun langkah-langkahnya adalah sebagai berikut

a Menghitung jarak antara dua amatan yang berskala nominal menggunakan MVDM dengan formula sebagai berikut

$\Delta(x,y) = w_x w_y \sum_{i=1}^{N} \delta(v_{1i}, v_{2i})$,

dimana $\Delta(x,y)$ adalah jarak antara amatan x dan y, w_x adalah bobot amatan x (dapat diabaikan), w_y adalah bobot amatan y (dapat diabaikan), N adalah banyaknya variabel penjelas Kemudian $\delta(v_{1i}v_{2i})$ adalah jarak antara amatan x dan y pada variabel ke-i dengan perhitungan jarak antar amatan x dan y pada variabel ke-i dilakukan melalui

$$\delta(v_{1i}, v_{2i}) = \sum_{j=1}^{S} \left| \frac{C_{1j}}{C_1} - \frac{C_{2j}}{C_2} \right|^k ,$$

dimana S adalah banyaknya kelas pada variabel respon, C_{xj} adalah banyaknya kategori x pada kelas ke-j, C_{yj} adalah banyaknya kategori y pada kelas ke-j, C_x adalah banyaknya kategori x terjadi, C_y adalah banyaknya kategori y yang terjadi dan k adalah konstanta (biasanya bernilai 1)

b Untuk nilai yang bersifat nominal, kategori mayoritas yang dipilih adalah kategori yang letaknya antara amatan kelas minoritas dengan k-tetangga terdekatnya Jika nilainya sama maka dipilih secara acak

c Nilai yang terpilih tersebut merupakan amatan yang baru

Dalam studi ini, teknik SMOTE dengan 350%, 600%, *over-sampling* 250% dan *under-sampling* 150% digunakan untuk menghasilkan tiga dataset baru

3 Hasil dan Pembahasan

Gambaran umum tentang variabel penjelas data obesitas di Indonesia pada Survei RISKESDAS 2013 disajikan pada Tabel 1 Pada Tabel 1, dari 722 329 status obesitas di Indonesia, sebanyak 101 023 (13,99%) orang mengalami obesitas, 611 557 (84,66%) orang tidak mengalami obesitas dan sebanyak 9 749 (1,32%) orang tidak diketahui status obesitasnya Pada Tabel 1 ini juga terlihat jumlah kelas obesitas dan tidak obesitas terlihat tidak seimbang dengan ketidakseimbangan kelas data berada pada kriteria sedang

Tabel 1 Gambaran umum tentang data status obesitas RISKESDAS 2013 di Indonesia

Status Obesitas	Frekuensi	Persentasi
Obesitas	101 023	13,99%
Tidak obesitas	611 557	84,66%
"NA"	9 749	1,32%
Jumlah	722 329	100%

Banyaknya variabel pada data akan menimbulkan masalah pada tahap pengolahan data. Dalam tahapan ini, seleksi variabel dilakukan dengan menggunakan metode *Chi-Square* (χ^2). Dari hasil seleksi variabel diketahui bahwa terdapat 11 variabel yang dipilih untuk di analisis (*p-value* < 0,01). Tabel 2 memperlihatkan karakteristik data status obesitas Survei RISKESDAS 2013 Indonesia sebelum dilakukan penyeimbangan data kelas berdasarkan variabel.

Tabel 2. Karakteristik Status obesitas RISKESDAS 2013 di Indonesia tanpa menggunakan teknik SMOTE

Variabel dan Atribut	Status obesitas		Total
	Obesitas	Tidak obesitas	
Jenis kelamin (X1)			
Laki-laki	31.724	311.270	342.994
Perempuan	67.391	294.592	361.983
Total	99.115	605.862	704.977
Umur (X2)			
15 - 34 Tahun	25.475	257.554	283.029
35 - 54 Tahun	56.830	231.286	288.116
55 - 74 Tahun	16.015	102.429	118.444
75 - 94 Tahun	789	14.357	15.146
95 - 114 Tahun	6	233	239
115 - 134 Tahun	0	3	3
Total	99.115	605.862	704.977
Merokok (X3)			
Ya	19.543	203.348	222.891
Tidak	79.572	402.514	482.086
Total	99.115	605.862	704.977
Aktifitas Berat (X4)			
Ya	29.905	246.243	276.148
Tidak	69.210	359.619	428.829
Total	99.115	605.862	704.977
Aktifitas Sedang (X5)			
Ya	88.970	518.483	607.453
Tidak	10.145	87.379	97.524
Total	99.115	605.862	704.977
Makan Buah (X6)			
Tidak Pernah	10.316	85.930	96.246
Satu Hari	19.247	143.388	162.635
Dua Hari	19.444	127.304	146.748
Tiga Hari	20.753	119.217	139.970
Empat Hari	7.409	41.661	49.070
Lima Hari	3.871	20.177	24.048
Enam Hari	1.630	7.988	9.618

Variabel dan Atribut	Status obesitas		Total
	Obesitas	Tidak obesitas	
Tujuh Hari	16 445	60 197	76 642
Total	99 115	605 862	704 977
Makan Sayuran (X7)			
Tidak Pernah	1 211	8 496	9 707
Satu Hari	2 967	21 273	24 240
Dua Hari	4 614	32 626	37 240
Tiga Hari	9 258	64 728	73 986
Empat Hari	6 795	48 720	55 515
Lima Hari	5 950	44 676	50 626
Enam Hari	5 638	41 661	47 299
Tujuh Hari	62 682	343 682	406 364
Total	99 115	605 862	704 977
Makanan Manis (X8)			
Kurang 1 kali per hari,	21 493	137 827	159 320
Sekali per hari,	30 645	180 807	211 452
3 sampai 6 kali per minggu,	16 332	105 750	122 082
1 sampai 2 kali per minggu,	18 617	114 371	132 988
Lebih dari 3 kali per bulan,	6 972	40 536	47 508
Tidak pernah	5 056	26 571	31 627
Total	99 115	605 862	704 977
Makanan Asin (X9)			
Kurang 1 kali per hari,	7 092	49 085	56 177
Sekali per hari,	12 687	83 881	96 568
3 sampai 6 kali per minggu,	17 513	3 619	131 132
1 sampai 2 kali per minggu,	29 693	185 251	214 944
Lebih 3 kali per bulan,	21 019	114 322	135 341
Tidak pernah	11 111	59 704	70 815
Total	99 115	605 862	704 977
Makanan Berlemak (X10)			
Kurang 1 kali per hari,	14 416	77 542	91 958
Sekali per hari,	22 307	120 775	143 082
3 sampai 6 kali per minggu,	23 162	140 589	163 751
1 sampai 2 kali per minggu,	26 786	172 629	199 415
Lebih 3 kali per bulan,	9 875	73 996	83 871
Tidak pernah	2 569	20 331	22 900
Total	99 115	605 862	704 977
Stres (X11)			
Tidak Stres	88 269	24 712	34 958
Stres	10 846	581 750	670 019
Total	99 115	605 862	704 977

Pada data asli obesitas RISKESDAS 2013 ini, data kelas minor adalah status obesitas, dan data kelas mayor adalah status tidak obesitas. Berdasarkan Tabel 1, data asli obesitas Survei RISKESDAS 2013 Indonesia yang mengalami ketidakseimbangan data kelas itu telah mengakibatkan amatan di kelas minor (obesitas) cenderung diabaikan dalam pengklasifikasian status obesitas. Oleh karena itu, permasalahan tersebut ditangani menggunakan teknik SMOTE dengan dua persentase berbeda yaitu 350% dan 600%. Hal ini berarti bahwa data dikelas minor (obesitas) akan dibangkitkan sebesar masing-masing 3,5 kali dan 6 kali.

Tahapan selanjutnya adalah melakukan penghitungan jarak amatan dengan menggunakan MVDM untuk memperoleh jarak amatan seperti yang telah dijelaskan di bagian 2.3. Data pada kelas minor (obesitas) dipilih secara acak untuk menentukan 5 tetangga terdekat. Kemudian, kategori mayoritas yang terdapat pada vektor utama dan 5 tetangganya untuk variabel bersangkutan akan dipilih. Jika terjadi nilai yang sama, maka dipilih secara acak, dan nilai yang terpilih tersebut merupakan data sampel buatan yang baru.

Pada studi ini, SMOTE digunakan pada kelas data yang memiliki atau mempunyai kelipatan yang sama. Jika tidak, maka dilakukan penambahan atau pengurangan (penghapusan) data secara acak pada kelas mayor, yang mendekati kelipatan kelas minor. Kelas obesitas dan kelas tidak obesitas mempunyai kelipatan tidak sama.

Pada SMOTE dengan 350% dibangkitkan data buatan sebesar 3,5 kali data minor sehingga diperoleh data baru sebanyak 346.902. Kemudian data baru ini ditambahkan pada data asli sehingga jumlah data baru yaitu 396.460. Jumlah data sebelum SMOTE adalah 99.115 dan setelah diterapkan SMOTE menjadi 396.460. Pada data hasil SMOTE dengan 350%, perbandingan antara kelas mayor dan kelas minor adalah 40% pada kelas minor (obesitas) dan 60% pada kelas mayor (tidak obesitas) (Tabel 3). Deskripsi data status obesitas RISKESDAS 2013 setelah diterapkan SMOTE dengan 350% dapat dilihat pada Tabel 4. Perbandingan kelas minor (obesitas) dan kelas mayor (tidak obesitas) dari setiap variabel tampak belum seimbang (Tabel 4).

Tabel 3. Hasil teknik SMOTE dengan 350% pada data obesitas RISKESDAS 2013

Status Obesitas	Tanpa SMOTE	Persentase SMOTE 350%
Obesitas	99.115 (14,05%)	396.460(40%)
Tidak obesitas	605.862 (85,95%)	594.690 (60%)
Jumlah	722.329 (100%)	991.150 (100%)

Tabel 4 Deskripsi data obesitas RISKESDAS 2013 Menggunakan SMOTE dengan 350%

Variabel dan Atribut	Status obesitas		Total
	Obesitas	Tidak obesitas	
Jenis kelamin (X1)			
Laki-laki	184 748	305 369	490 117
Perempuan	211 712	289 321	501 033
Total	396 460	594 690	991 150
Umur (X2)			
15 - 34 Tahun	120 136	252 139	372 275
35 - 54 Tahun	188 971	227 390	416 361
55 - 74 Tahun	82 854	100 967	183 821
75 - 94 Tahun	4 486	13 974	18 460
95 - 114 Tahun	13	218	231
115 - 134 Tahun	0	2	2
Total	396 460	594 690	991 150
Merokok (X3)			
Ya	165 238	199 081	364 319
Tidak	231 222	395 609	626 831
Total	396 460	594 690	991 150
Aktifitas berat (X4)			
Ya	177 219	241 373	418 592
Tidak	219 241	353 317	572 558
Total	396 460	594 690	991 150
Aktifitas sedang (X5)			
Ya	239 502	508 987	748 489
Tidak	156 958	85 703	242 661
Total	396 460	594 690	991 150
Makan buah (X6)			
Tidak Pernah	52 964	84 826	137 790
Satu Hari	84 147	140 505	224 652
Dua Hari	78 953	125 458	204 411
Tiga Hari	77 622	116 377	193 999
Empat Hari	30 108	40 647	70 755
Lima Hari	14 779	19 827	34 606
Enam Hari	6 135	7 962	14 097
Tujuh Hari	51 752	59 088	11 0840
Total	396 460	594 690	991 150
Makan sayuran (X7)			
Tidak Pernah	15 251	84 25	23 676
Satu Hari	18 713	20 916	39 629
Dua Hari	27 379	31 832	59 211
Tiga Hari	58 663	63 334	121 997

Variabel dan Atribut	Status obesitas		Total
	Obesitas	Tidak obesitas	
Empat Hari	35 330	47 977	83 307
Lima Hari	26 080	44 004	70 084
Enam Hari	27 803	41 125	68 928
Tujuh Hari	187 241	337 077	524 318
Total	396 460	594 690	991 150
Makanan manis (X8)			
Kurang 1 kali per hari,	93 296	134 660	227 956
Sekali per hari,	119 677	177 923	297 600
3 sampai 6 kali per minggu,	57 229	103 604	160 833
1 sampai 2 kali per minggu,	73 579	112 520	186 099
Lebih 3 kali per bulan,	27 628	39 882	67 510
Tidak pernah	25 051	26 101	51 152
Total	396 460	594 690	991 150
Makanan asin (X9)			
Kurang 1 kali per hari,	33 300	47 890	81 190
Sekali per hari,	49 129	82 058	131 187
3 sampai 6 kali per minggu,	70 098	112 010	182 108
1 sampai 2 kali per minggu,	109 483	181 311	290 794
Lebih 3 kali per bulan,	80 438	112 543	192 981
Tidak pernah	54 012	58 878	112 890
Total	396 460	594 690	991 150
Makanan berlemak (X10)			
Kurang 1 kali per hari,	60 415	76 079	136 494
Sekali per hari,	85 774	117 994	203 768
3 sampai 6 kali per minggu,	88 824	138 386	227 210
1 sampai 2 kali per minggu,	100 039	169 385	269 424
Lebih 3 kali per bulan,	46 177	72 879	119 056
Tidak pernah	15 231	19 967	35 198
Total	396 460	594 690	991 150
Stress (X11)			
Stres	158 499	23 697	182 196
Tidak Stres	237 961	570 993	808 954
Total	396 460	594 690	991 150

Selanjutnya penerapan teknik SMOTE dilakukan dengan menaikkan persentase ke *over-sampling* 600%. Data buatan dibangkitkan sebesar 6 kali data minor menjadi 594 690, sehingga diperoleh data baru sebanyak 693 805. Jumlah data sebelum dan sesudah diterapkan SMOTE dengan 600% adalah masing-masing 99 115 dan 693 805. Pada data hasil SMOTE dengan 600% ini, perbandingan antara kelas minor (obesitas) dan kelas mayor (tidak obesitas) menjadi nampak lebih seimbang yaitu 53,85% pada kelas obesitas dan 46,15% pada kelas tidak obesitas. Hasil teknik

SMOTE dengan kenaikan persentase sampai 600% dapat dilihat pada Tabel 5. Sebagai akibat, perbandingan kelas minor (obesitas) dan kelas mayor (tidak obesitas) dari setiap variabel terlihat seimbang (Tabel 6). Pada Tabel 5 disajikan gambaran data obesitas survei RISKESDAS 2013 Indonesia setelah dilakukan penyeimbangan data kelas berdasarkan variabel.

Tabel 5. Hasil teknik SMOTE dengan 600% pada data obesitas RISKESDAS 2013

Status Obesitas	Tanpa SMOTE	SMOTE 600%
Obesitas	99 115 (14,05%)	693 805(53,84%)
Tidak obesitas	605 862 (85,95%)	594 690 (46,16%)
Jumlah	722 329 (100%)	1 288 495 (100%)

Tabel 6. Deskripsi data obesitas RISKESDAS 2013 Menggunakan SMOTE dengan 600%

Variabel dan Atribut	Status obesitas		Total
	Obesitas	Tidak obesitas	
Jenis kelamin (X1)			
Laki-laki	335 315	305 448	640 763
Perempuan	358 490	289 242	647 732
Total	693 805	594 690	1 288 495
Umur (X2)			
15 - 34 Tahun	187 693	252 954	440 647
35 - 54 Tahun	316 288	227 106	543 394
55 - 74 Tahun	177 759	100 367	278 126
75 - 94 Tahun	12 040	14 025	26 065
95 - 114 Tahun	25	237	262
115 - 134 Tahun	0	1	1
Total	693 805	594 690	128 8495
Merokok (X3)			
Ya	305 957	199 115	505 072
Tidak	387 848	395 575	783 423
Total	693 805	594 690	1 288 495
Aktifitas berat (X4)			
Ya	316 295	241 303	557 598
Tidak	377 510	353 387	730 897
Total	693 805	594 690	1 288 495
Aktifitas sedang (X5)			
Ya	395 439	509 174	904 613
Tidak	298 366	85 516	383 882
Total	693 805	594 690	1 288 495

Makan buah (X6)			
Tidak Pernah	114 363	84 440	198 803
Satu Hari	147 474	140 498	287 972
Dua Hari	133 919	125 150	259 069
Tiga Hari	132 543	116 989	249 532
Empat Hari	43 621	40 494	84 115
Lima Hari	24 695	19 811	44 506
Enam Hari	8 194	7 900	16 094
Tujuh Hari	88 996	59 408	148 404
Total	693 805	594 690	1 288 495
Makan sayuran (X7)			
Tidak Pernah	33 323	8 233	41 556
Satu Hari	37 393	20 910	58 303
Dua Hari	39 335	31 969	71 304
Tiga Hari	122 558	63 891	186 449
Empat Hari	64 058	48 056	112 114
Lima Hari	40 333	43 814	84 147
Enam Hari	43 566	41 092	84 658
Tujuh Hari	313 239	336 725	649 964
Total	693 805	594 690	1 288 495
Makanan manis (X8)			
Kurang 1 kali per hari,	144 767	135 812	280 579
1 kali per hari,	190 085	177 015	367 100
3 sampai 6 kali per minggu,	95 066	104 132	199 198
1 sampai 2 kali per minggu,	149 148	111 596	260 744
Lebih 3 kali per bulan,	61 202	39 969	101 171
Tidak pernah	53 537	26 166	79 703
Total	693 805	594 690	1 288 495
Makanan asin (X9)			
Kurang 1 kali per hari,	47 452	48 041	95 493
Sekali per hari,	91 010	82 599	173 609
3 sampai 6 kali per minggu,	119 109	111 737	230 846
1 sampai 2 kali per minggu,	186 465	181 915	368 380
Lebih 3 kali per bulan,	139 576	112 015	251 591
Tidak pernah	110 193	58 383	168 576
Total	693 805	594 690	1 288 495
Makanan berleman (X10)			
Kurang 1 kali per hari,	91 076	75 953	167 029
Sekali per hari,	155 704	119 164	274 868
3 sampai 6 kali per minggu,	147 277	137 753	285 030
1 sampai 2 kali per minggu,	186 841	169 408	356 249

Lebih 3 kali per bulan,	87 023	72 512	159 535
Tidak pernah	25 884	19 900	45 784
Total	693 805	594 690	1 288 495
Stres (X11)			
Tidak Stres	306 089	23 758	329 847
Stres	387 716	570 932	958 648
Total	693 805	594 690	1 288 495

Tabel 7 Hasil teknik SMOTE dengan *over-sampling* 250% dan *under-sampling* 150% pada data obesitas RISKESDAS 2013

Status Obesitas	Tanpa SMOTE (%)	SMOTE *over-sampling* 250% dan *under-sampling* 150% (%)
Obesitas	99 115 (14,05)	297 345 (50,00)
Tidak obesitas	605 862 (85,95)	297 345 (50,00)
Jumlah	722 329 (100,00)	594 690 (100,00)

Modifikasi teknik SMOTE dengan kombinasi *over-sampling* dan *under-sampling* juga dilakukan yaitu *over-sampling* 250% dan *under-sampling* 150%. Pada *over-sampling* 250% dibangkitkan data buatan sebesar 2,5 kali data kelas minor sehingga diperoleh 247 787,5 data baru. Kemudian data baru ini ditambahkan pada data asli yang hasilnya sebanyak 297 345. Jumlah data sebelum dan sesudah diterapkan SMOTE adalah masing 99 115 dan 297 345. Selanjutnya untuk *under-sampling* 150%, data buatan dihapus sebesar 1.5 kali data mayor sehingga diperoleh 396 460 data baru. Data baru ini kemudian dikurangkan dengan jumlah data kelas minor sehingga diperoleh 297 345. Jumlah data sebelum dan sesudah diterapkan SMOTE adalah masing-masing 594 690 dan 297 345. Pada data hasil SMOTE ini, perbandingan antara kelas minor (obesitas) dan kelas mayor (tidak obesitas) menjadi lebih seimbang yaitu 50% pada kelas obesitas dan 50% pada kelas mayor (tidak obesitas). Hasil teknik SMOTE dengan kombinasi *over-sampling* (250%) dan *under-sampling* (150%) ini dapat dilihat pada Tabel 7. Tabel 8 menyajikan gambaran data obesitas Survei RISKESDAS 2013 Indonesia setelah dilakukan penyeimbangan data kelas berdasarkan variabel.

Dari Tabel 3, 5 dan 7 dapat dibandingkan jumlah kejadian obesitas dan tidak obesitas berdasarkan persentase teknik SMOTE (Tabel 9). Dari Tabel 9 diketahui bahwa teknik SMOTE dengan 600% dapat meningkatkan data kelas minor (obesitas) dengan lebih baik dan perbandingan kelas minor (obesitas) dan kelas mayor (tidak obesitas) menjadi mendekati seimbang. Sementara itu, teknik kombinasi *over-sampling* dan *under-sampling* membuat dataset menjadi seimbang.

Tabel 8 Deskripsi data obesitas RISKESDAS 2013 Menggunakan SMOTE dengan *over-sampling* 250% dan *under-sampling* 150%

Variabel dan Atribut	Status obesitas		Total
	Obesitas	Tidak obesitas	
Jenis kelamin (X1)			
Laki-laki	134 150	152 599	286 749
Perempuan	163 195	144 746	307 941
Total	297 345	297 345	594 690
Umur (X2)			
15 - 34 Tahun	88 321	126 689	215 010
35 - 54 Tahun	144 902	113 778	258 680
55 - 74 Tahun	60 846	49 767	110 613
75 - 94 Tahun	3 264	7 004	10 268
95 - 114 Tahun	12	104	116
115 - 134 Tahun	0	3	3
Total	297 345	297 345	594 690
Merokok (X3)			
Ya	116 378	99 798	216 176
Tidak	180 967	197 547	378514
Total	297 345	297 345	594 690
Aktifitas berat (X4)			
Ya	128 258	120 694	248 952
Tidak	169 087	176 651	345 738
Total	297 345	297 345	594 690
Aktifitas sedang (X5)			
Ya	189 062	254 418	443 480
Tidak	108 283	42 927	151 210
Total	297 345	297 345	594 690
Makan Buah (X6)			
Tidak Pernah	38 869	42 513	81 382
Satu Hari	62 347	70 158	132 505
Dua Hari	58 926	62 318	121 244
Tiga Hari	58 555	58 862	117 417
Empat Hari	22 591	20 228	42 819
Lima Hari	11 279	9 905	21 184
Enam Hari	4 561	3 945	8 506
Tujuh Hari	40 217	29 416	696 33
Total	297 345	297 345	594 690
Makan sayuran (X7)			
Tidak Pernah	10 481	4 178	14 659
Satu Hari	13 543	10 482	24 025
Dua Hari	19 881	15 944	35 825
Tiga Hari	42 164	31 865	74 029

Variabel dan Atribut	Status obesitas		Total
	Obesitas	Tidak obesitas	
Empat Hari	25 894	23 887	49 781
Lima Hari	19 505	22 007	41 512
Enam Hari	20 213	20 484	40 697
Tujuh Hari	145 664	168 498	314 162
Total	297 345	297 345	594 690
Makanan manis (X8)			
Kurang 1 kali per hari,	69 214	67 416	136 630
Sekali per hari,	89 943	88 838	178 781
3 sampai 6 kali per minggu,	43 620	52 056	95 676
1 sampai 2 kali per minggu,	55 583	55 959	111 542
Lebih 3 kali per bulan,	20 678	19 886	40 564
Tidak pernah	18 307	13 190	31 497
Total	297 345	297 345	594 690
Makanan asin (X9)			
Kurang 1 kali per hari,	24 675	24 043	48 718
Sekali per hari,	37 044	41 306	78 350
3 sampai 6 kali per minggu,	52 626	55 740	108 366
1 sampai 2 kali per minggu,	82 729	90 731	173 460
Lebih 3 kali per bulan,	60 456	56 378	116 834
Tidak pernah	39 815	29 147	68 962
Total	297 345	297 345	594 690
Makanan berlemak (X10)			
Kurang 1 kali per hari,	45 043	38 097	83 140
Sekali per hari,	64 436	58 973	123 409
3 sampai 6 kali per minggu,	66 766	69 247	136 013
1 sampai 2 kali per minggu,	76 011	84 965	160 976
Lebih 3 kali per bulan,	34 096	36 153	70 249
Tidak pernah	10 993	9 910	20 903
Total	297 345	297 345	594 690
Stres (X11)			
Stres	109 116	11 901	121 017
Tidak Stres	188 229	285 444	473 673
Total	297 345	297 345	594 690

Tabel 9 Jumlah kejadian obesitas dan tidak obesitas berdasarkan persentase teknik SMOTE

Persentase Teknik sampling (%)	Kelas Tidak obesitas	Kelas Obesitas
SMOTE Over-sampling 350	594 690	396 460
SMOTE Over-sampling 600	594 690	693 805
Over-sampling 250%, *under sampling* 150%	297 345	297 345

4 Simpulan

Eksplorasi status obesitas dari survei RISKESDAS 2013 Indonesia telah dilakukan dengan menggunakan teknik tanpa SMOTE, teknik SMOTE dan teknik kombinasi *over-sampling* dan *under-sampling* Hasilnya menunjukkan jumlah kelas obesitas dan tidak obesitas mengalami ketidakseimbangan kelas data dengan kriteria sedang Ketidakseimbangan kelas data ini telah di atasi dengan teknik SMOTE dan teknik *over-sampling* dan *under-sampling* Pada penggunaan teknik *over-sampling* dan *under-sampling* dataset kelas obesitas dapat dibuat seimbang dengan mudah Namun, *over-sampling* pada dataset kelas minoritas (obesitas) menuju pada hasil yang *overfitting* Sementara teknik SMOTE dengan 600% dapat meningkatkan data kelas minor (obesitas) dengan lebih baik Dalam studi yang akan datang, hasil yang telah diperoleh ini akan divalidasi dengan metode mengatasi ketidakseimbangan data lainnya dan prediksi kinerja metode klasifikasi pembelajaran mesin

Ucapan Terima Kasih Penulis pertama berterima kasih kepada Kementerian Riset dan Teknologi/ Badan Riset dan Inovasi Nasional yang telah membiayai studi ini melalui Skema PDUPT tahun anggaran 2020 nomor kontrak 1516/UN4 22/PT 01 03/2020 Selain itu penulis pertama juga mengucapkan terima kasih kepada Kementerian Kesehatan melalui Badan Penelitian dan Pengembangan Masyarakat yang telah memberikan akses ke data RISKESDAS Indonesia

Daftar Pustaka

Alghamdi, M , Al-Mallah, M , Keteyian, S , Brawner, C , Ehrman, J , & Sakr, S (2017) Predicting diabetes mellitus using SMOTE and ensemble machine learning approach The Henry Ford Exercise Testing (FIT) project *PLOS ONE*, *12*(7) e0179805

ASEAN / UNICEF / WHO Regional Report (2016) *World health statistics 2016 monitoring health for the SDGs, sustainable development goals*

Branco, P , Torgo, L , & Ribeiro, R P (2016) A Survey of Predictive Modeling on Imbalanced Domains *ACM Comput Surv* , *49*(2) https //doi org/10 1145/2907070

Chawla, N V (2005) Data Mining for Imbalanced Datasets An Overview In O Maimon & L Rokach (Eds), *Data Mining and Knowledge Discovery Handbook* (pp 853–867) https //doi org/10 1007/0-387-25465-X_40

Chawla, N V , Bowyer, K W , Hall, L O , & Kegelmeyer, W P (2002) SMOTE Synthetic Minority Over-sampling Technique *Journal of Artificial Intelligence Research*, *16* 321–357 https //doi org/10 1613/jair 953

Chen, X , & Wasikowski, M (2008) FAST A Roc-Based Feature Selection Metric for Small Samples and Imbalanced Data Classification Problems *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 124–132 https //doi org/10 1145/1401890 1401910

Cost, S , & Salzberg, S (1993) A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features *Machine Learning*, *10*(1) 57–78 https //doi org/10 1023/A 1022664626993

Ente, D R , Thamrin, S A , Arifin, S , Kuswanto, H , & Andreza, A (2020) Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4 5 *Indonesian Journal of Statistics and Its Applications*, *4*(1) 80–88 https //doi org/10 29244/ijsa v4i1 330

Jo, T , & Japkowicz, N (2004) Class Imbalances versus Small Disjuncts *SIGKDD Explor Newsl* , *6*(1) 40–49 https //doi org/10 1145/1007730 1007737

[KEMENKES RI], Kementerian Kesehatan Republik Indonesia (2014) *Profil Kesehatan Indonesia Tahun 2013* Jakarta (ID) Kemenkes RI

Krawczyk, B (2016) Learning from imbalanced data open challenges and future directions *Progress in Artificial Intelligence*, *5*(4) 221–232 https //doi org/10 1007/s13748-016-0094-0

Mustaqim, M , Warsito, B , & Surarso, B (2019) Kombinasi Synthetic Minority Oversampling Technique (SMOTE) dan Neural Network Backpropagation untuk menangani data tidak seimbang pada prediksi pemakaian alat kontrasepsi implan *Register Jurnal Ilmiah Teknologi Sistem Informasi*, *5*(2) 116–127 https //doi org/10 26594/register v5i2 1705

[RISKESDAS], Riset Kesehatan Dasar (2013) *Hasil Riset Kesehatan Dasar 2013* Jakarta (ID) Kemenkes RI

Selya, A S , & Anshutz, D (2018) Machine learning for the classification of obesity from dietary and physical activity patterns In *Advanced Data Analytics in Health* (pp 77–97) Springer

Sun, Y , Wong, A K C , & Kamel, M S (2009) Classification of Imbalanced data a review *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(4) 687–719 https //doi org/10 1142/S0218001409007326

Witten, I H , Frank, E , & Hall, M A (2011) *Data mining practical machine learning tools and techniques* (3rd ed) https //doi org/10 1016/C2009-0-19715-5

Yap, B W , Rani, K A , Rahman, H A A , Fong, S , Khairudin, Z , & Abdullah, N N (2014) An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets In T Herawan, M Deris, & J Abawajy (Eds), *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)* (pp 13–22) https //doi org/10 1007/978-981-4585-18-7_2

Zhang, D , Liu, W , Gong, X , & Jin, H (2011) A Novel Improved SMOTE Resampling Algorithm Based on Fractal *Computational Information Systems*, 2204–2211

Zhu, T , Lin, Y , & Liu, Y (2017) Synthetic Minority Oversampling Technique for Multiclass Imbalance Problems *Pattern Recogn* , *72*(C) 327–340 https //doi org/10 1016/j patcog 2017 07 024