# COMBINE UNDERSAMPLING UNTUK MENANGANI DATA TIDAK SEIMBANG PADA LAMA BELAJAR SISWA DI RUMAH

Rifaldi Ardiansyah Alqaida*, Gusti Ngurah Adhi Wibawa, Irma Yahya, Bahridin Abapihi, Rasdiyanah, Lilis Laome, Ruslan
*Program Studi S1 Statistik,FMIPA, Universitas Halu Oleo*
*Email: rifaldiardiansyah14@gmail.com*

***ABSTRACT***

*Due to the Covid-19 pandemic, students were forced to stay home and study online. After conducting a survey in six districts in Southeast Sulawesi province, the data was obtained was not balanced between students who studied ≥3 hours and < 3 hours with ratio of 86% (974 : 14% (164) so it was necessary to apply the resampling method to obtain appropriate conclusions. The objectives to be achieved in this study are (1) to determine the results of unbalanced data classification by logistic regression using the tomek link SMOTE method and combined undersampling. (2) determine the factors that influence the length of time students study at home. The results of the analysis show that when viewed from the sensitivity value produced by the combined undersampling technique is the highest (78.26) where the significant variable is Variable X3 (whether students like online learning or not).*

***Keywords: Student learning time at home***

**ABSTRAK**

Akibat pandemi Covid 19 siswa sekolah terpaksa dirumahkan dan belajar secara daring. Setelah dilakukan survey pada enam kabupaten di provinsi Sulawesi Tenggara diperoleh data yang tidak seimbang antara siswa yang belajar ≥ 3 jam dan < 3 jam dengan rasio kelas 86% (974) : 14% (164) sehingga perlu diterapkan metode resampling demi memperoleh kesimpulan yang sesuai. Tujuan yang ingin dicapai dalam penelitian ini adalah (1) untuk mengetahui hasil klasifikasi data tidak seimbang dengan regresi logistik menggunakan metode SMOTE tomek link dan combined undersampling. (2) menentukan faktor-faktor yang mempengaruhi lama siswa belajar di rumah. Hasil analisis menunjukkan bahwa jika ditinjau dari nilai sensitivitas yang dihasilkan teknik resampling combine undersampling adalah yang paling tinggi dengan nilai 78,26 dimana variabel signifikan adalah Variabel X3 (apakah siswa suka atau tidak terhadap pembelajaran daring).

**Kata Kunci: Lama belajar Siswa di rumah**

## PENDAHULUAN

Belajar merupakan kunci utama dari kesuksesan siswa dalam pendidikan. Dari proses belajar yang dilakukan siswa dapat mengetahui apa yang belum diketahui dan memperdalam apa yang sudah diketahui baik belajar yang dilakukan di sekolah maupun di rumah. Untuk memastikan siswa belajar secara mandiri di rumah, diperlukan kerjasama yang baik antara guru dan orang tua siswa karena tanggung jawab pendidikan anak tidak hanya menjadi tanggung jawab guru, tetapi juga merupakan tanggung jawab orang tua. Hal ini sangat penting untuk dilakukan agar siswa tetap dapat belajar secara maksimal sehingga hasil belajarnya juga maksimal.

Faktor keluarga merupakan salah satu faktor yang sangat berpengaruh terhadap belajar anak, yang meliputi: a) cara orang tua mendidik, b) relasi antara anggota keluarga, c) suasana rumah tangga, d) keadaan ekonomi keluarga, e) latar belakang kebudayaan (Slameto, 2003). Belajar yang dilakukan di rumah merupakan kegiatan setelah pembelajaran yang dilakukan siswa di sekolah dengan cara mengulang pelajaran yang telah diajarkan, sangat membantu dan menunjang pemahaman materi siswa. Hal ini bertujuan untuk mengingat kembali pelajaran yang telah diajarkan dan memperbaiki semua kesan yang masih samar-samar untuk menjadi kesan yang sesungguhnya yang tergambar jelas dalam ingatan. Maksudnya, materi yang sudah dipelajari siswa di sekolah akan lebih jelas dan paham apabila dipelajari ulang dan apabila materi yang diajarkan guru di sekolah siswa belum paham, dengan belajar di rumah melalui cara

mengulang pelajaran tadi diharapkan siswa bisa jelas dan paham.

Belajar yang dilakukan di rumah juga dapat dilakukan untuk mempersiapkan pelajaran yang akan datang/pokok bahasan baru sebagai pemahaman awal untuk materi pelajaran yang akan diajarkan selanjutnya.Lama belajar merupakan waktu yang dihabiskan anak dalam proses belajar yang dilakukan yang diakumulasi menjadi satu dalam satu hari atau dalam satu kali belajar. Lama belajar yang dilakukan anak dalam satu hari atau dalam satu kali belajar berbeda-beda dari anak satu dengan yang lainnya. Berdasarkan hasil survey yang telah dilakukan di enam kabupaten di Sulawesi tenggara baik dari orang tua maupun anak mereka, diperoleh sejumlah data berupa lama belajar siswa dan beberapa indikator yang diperkirakan memiliki pengaruh pada lama waktu belajar dari para siswa tersebut apakah mereka meluangkan waktunya minimal selama 3 jam perhari untuk belajar di rumah atau kurang dari itu. Dari data yang ada diperoleh situasi data yang tidak seimbang atau imbalance antara siswa yang belajar minimal 3 jam dan yang belajar kurang dari 3 jam.

Setelah dilakukan survey pada enam kabupaten di provinsi Sulawesi Tenggara diperoleh data yang tidak seimbang antara siswa yang belajar $\geq$ 3 jam dan $<$ 3 jam dengan rasio kelas 86% (974) : 14% (164). Apriana (2016) mengatakan bahwa salah satu permasalahan dalam klasifikasi data adalah komposisi data yang tidak seimbang (imbalanced data). Salah satu permasalahan dalam klasifikasi data adalah komposisi data yang tidak seimbang (Apriana, 2016). Salah satu cara yang umum digunakan untuk menyelesaikan masalah data tidak seimbang yaitu dengan cara menyeimbangkan data yang dapat dilakukan dengan 2 teknik resampling antara lain undersampling, yaitu mengurangi kelas mayoritas agar jumlahnya sebanding dengan kelas minoritas dan oversampling dengan menambah data kelas minoritas hingga sebanding dengan data kelas mayoritas Menurut (Faisal & Nugrahadi, 2016). Tomek link merupakan metode undersampling hasil pengembangan dari metode Condensed Nearest Neighbor Rule (CNN) (Hart, 1967). Berbeda dengan CNN, tujuan utama tomek link yaitu menghapus noise dan borderline pada data yang dapat mempersulit proses klasifikasi pada data tidak seimbang. SMOTE (Synthetic Minority Oversampling Technique) merupakan salah satu metode dalam penanggulangan data tidak seimbang yang diusulkan oleh (Chawla dkk, 2002).

Random Undersampling (RUS) menghitung selisih antara kelas mayoritas dan minoritas kemudian dilakukan perulangan. Selisih hasil perhitungan, selama perulangan data kelas mayoritas dihapus secara acak, sehingga jumlah kelas mayoritas sama dengan minoritas. (Yu dkk, 2017). Mengkombinasikan random undersampling dan tomek link untuk menangani data tidak seimbang dapat digunakan untuk memperoleh hasil klasifikasi yang lebih baik (Shintia, 2018). Analisis regresi logistik biner digunakan untuk menggambarkan hubungan antara variabel respon dengan sekumpulan variabel prediktor, dimana variabel respon bersifat biner atau dikotomus yang hanya mempunyai dua kemungkinan nilai misalnya sukses dan gagal (Misna dkk, 2018).

Berdasarkan latar belakang diatas, penulis tertarik untuk mengadakan penelitian tentang SMOTE Tomek Link dan Combine Undersampling Untuk Menangani Data Tidak Seimbang Pada Pemodelan Lama Belajar Di Rumah.

## METODE PENELITIAN

Survey yang dilakukan pada penelitian ini dilaksanakan pada enam kabupaten yang ada di provinsi Sulawesi tenggara. Hal yang pertama dilakukan setelah memproleh data adalah mendeskripsikan data dengan membuat diagram untuk melihat perbedaan kelas mayoritas dan kelas minoritas yang kemudian dilanjutkan dengan melakukan pembagian data menjadi data latih dan data uji menggunakan aplikasi R sebanyak 10 kali, lalu dilakukan penerapan teknik resampling pada tiap data latih dan dimodelkan dengan regresi logistik untuk kemudian dipilih ulangan dengan model yang paling

sering muncul. Setelah itu kemudian diterapkanlah teknik resampling SMOTE dan *undersampling* pada data latih ulangan terpilih.

Pada penerapan teknik SMOTE dan *undersampling* terlebih dahulu diterapkan metode tomek *link* guna mengeliminasi data yang bersifat nois pada data latih yang mana teknik ini didasarkan pada rumus k- nearest neighbor (KNN) untuk menentukan apakah suatu data adalah nois atau tomek *link* dan harus dihapus atau tidak.

$$d=\sqrt{(\boldsymbol{x_1}-\boldsymbol{y_1})^2+(\boldsymbol{x_2}-\boldsymbol{y_2})^2}$$

suatu data dikatakan nois atau tomek *link* jika tidak ada sampel $z$, sehingga $d(x,z) < d(x,y)$ dan $d(y,z) < d(x,y)$. Data latih kemudian diberi perlakuan SMOTE yang dalam proses pembangkitan datanya didasarkan pada metode KNN seperti pada kasus tomek *link.*

$$d(X,Z)=\sqrt{(x_1-z_1)^2+(x_2-z_2)^2+\ldots+(x_p-z_p)^2}$$

Untuk kemudian dilakukan pembentukan data sintetik

$$\boldsymbol{X_{syn}} = \boldsymbol{X_i} + (\boldsymbol{X_{knn}} - \boldsymbol{X_i})\boldsymbol{\gamma}$$

Inilah asal dari metode MOTE tomek *link* yang juga diterapkan pada teknik *undersampling* hingga menjadi *combine undersampling*.

Setelah menerapkan resampling pada data latih maka dilakukan pemodelan menggunakan regresi logistik pada data latih, data SMOTE tomek *link* dan data *combine undersampling*

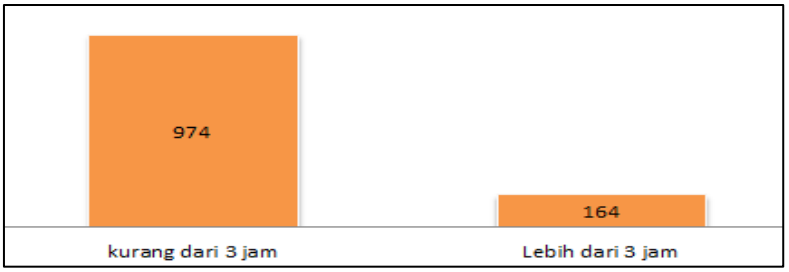
$$P(Y=1)=\pi(x)=\frac{\exp(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p)}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p)}$$

kemudian dilakukan uji simultan dan parsial pada tiap model untuk memperoleh model akhir.

Dari model akhir tersebut kemudian dibuat confusion matrix dengan menerapkan tiap model untuk mempredisi data uji. Dari confusion matrix kemudian diukur performa klasifikasi yang diperoleh dari tiap model berupa akurasi, sensitivitas, spesifisitas, G-mean dan AUC. Setelah itu, diterapkan teknik resampling yang terbaik pada data asli untuk melihat performa sebenarnya dari teknik resampling tersebut.

## HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini adalah data primer yang diperoleh dari hasil survey yang pernah dilakukan sebelumnya pada beberapa sekolah dasar yang tersebar di enam kabupaten di provinsi Sulawesi tenggara. Adapun kabupaten yang dimaksud yaitu Bombana, Buton Utara, Kolaka Timur, Konawe, Konawe Utara dan kabupaten Muna Barat. Data yang diperoleh merupakan data yang tidak seimbang dengan rasio kelas 86% (974) : 14% (164).




Gambar 1. Komposisi data

Dari Gambar 1 dapat dilihat bahwa Data yang diperoleh merupakan data yang

tidak seimbang dengan rasio kelas 974:164. Data kemudian dibagi menjadi dua sebagai data latih dan data uji dengan jumlah partisi data 90%:10%. Pembagian data dilakukan sebanyak 10 kali demi menemukan model yang paling sering muncul.

Tabel 1. Model Akhir pada Tiap Pembagian Data

| Pembagian data | Tanpa *resampling* | *Combine undersampling* | SMOTE tomek *link* |
|---|---|---|---|
| 1 | $X_1$+$X_3$ | $X_3$ | $X_1$ +$X_3$ |
| 2 | $X_1$ +$X_3$ | $X_3$ | $X_1$ +$X_3$ |
| 3 | $X_1$+$X_3$ | $X_3$ | $X_1$ +$X_3$ |
| 4 | $X_3$ | $X_1$ +$X_6$ | $X_1$ +$X_3$ |
| 5 | $X_1$ +$X_3$ | $X_1$ +$X_3$ | $X_1$ +$X_3$ |
| 6 | $X_3$ | $X_3$ | $X_1$ +$X_3$ |
| 7 | $X_3$ | $X_3$ | $X_1$ +$X_3$+$X_5$ |
| 8 | $X_3$ | $X_1$ +$X_3$ | $X_1$ +$X_3$ |
| 9 | $X_3$ | $X_1$ +$X_3$ | $X_1$ +$X_3$ |
| 10 | $X_3$ | $X_1$ +$X_3$ | $X_1$ +$X_3$+$X_5$ |

Dari Tabel 1 dapat dilihat bahwa untuk tanpa resampling model akhir yang paling sering muncul adalah Y~X3. Pada Combine undersampling model akhir yang paling sering muncul adalah Y~X3 dan pada SMOTE tomek link model akhir yang paling sering muncul adalah Y~X1+X3. Oleh karena itu maka dipilihlah pembagian data ke-6 untuk melakukan perbandingan klasifikasi dikarenakan model akhir untuk tiap resamplingnya sesuai dengan model akhir yang paling sering muncul baik pada tanpa resampling, combine undersampling maupun model akhir SMOTE tomek link.

Tabel 2. Jumlah Pengamatan Partisi Data 90%:10%

| Kelas | Data asli | Tanpa *resampling* | | *Combined undersampling* | | SMOTE tomek *link* | |
|---|---|---|---|---|---|---|---|
| | | Latih | Uji | Latih | Uji | Latih | Uji |
| Minoritas | 164 | 141 | 16 | 141 | 16 | 847 | 16 |
| Mayoritas | 974 | 884 | 97 | 141 | 97 | 883 | 97 |
| Total | 1138 | 1025 | 113 | 282 | 113 | 1730 | 113 |

Setelah membagi data menjadi data latih dan data uji yang kemudian diklasifikasi secara tanpa resampling, combined undersampling dan SMOTE tomek link, dilakukanlah analisis regresi logistik terhadap data hasil klasifikasi dari ketiga metode klasifikasi tersebut.

Tabel 3. Model Awal

| | Tanpa *resampling* | *Combine undersampling* | SMOTE tomek *link* |
|---|---|---|---|
| *Intercept* | -17,478 | 0,058 | 0,022 |
| $X_1$ | -0,277 | -0.468 | -0,328 |
| $X_2$ | 0,0003 | 0,114 | 0,075 |
| $X_3$ | -0,955 | -1,190 | -1,363 |
| $X_4$ | -0,068 | -0,098 | -0,054 |
| $X_5$ | 0,060 | 0,001 | 0,092 |
| $X_6$ | 0,206 | 0,368 | 0,130 |

Dari Tabel 3 dapat diperoleh model regresi logistik sebagai berikut:

Model regresi logistik data tanpa *resampling* :

$$P(Y=1)=\pi(x)=\frac{\exp(-1,748-0,277x_1+0,0003x_2-0,955x_3-0,068x_4+0,06x_5+0,206x_6)}{1+\exp(-1,748-0,277x_1+0,0003x_2-0,955x_3-0,068x_4+0,06x_5+0,206x_6)}$$

Model regresi logistik *combine undersampling* :

$$P(Y=1)=\pi(x)=\frac{\exp(0,058-0,468x_1+0,114x_2-1,19x_3-0,098x_4+0,001x_5+0,368x_6)}{1+\exp(0,058-0,468x_1+0,114x_2-1,19x_3-0,098x_4+0,001x_5+0,368x_6)}$$

Model regresi logistik SMOTE tomek *link* :

$$P(Y=1)=\pi(x)=\frac{\exp(0,022-0,328x_1+0,075x_2-1,363x_3-0,054x_4+0,092x_5+0,13x_6)}{1+\exp(0,022-0,328x_1+0,075x_2-1,363x_3-0,054x_4+0,092x_5+0,13x_6)}$$

Tabel 4. Signifikansi Parameter dengan Uji Rasio Likelihood

| Metode | $X^2_{tabel}$ | Df | $G^2$ |
|---|---|---|---|
| Tanpa Resampling | 12.592 | 6 | 20,077 |
| Combined Undersampling | 12.592 | 6 | 21,763 |
| SMOTE Tomek Link | 12.592 | 6 | 110,637 |

Dari Tabel 4 dapat dilihat bahwa hasil uji serentak model dari tanpa *resampling*, *combine undersampling* dan teknik SMOTE tomek *link* memiliki nilai $G^2$ lebih besar dari nilai $X^2_{tabel}$ =12.592 yang berarti menolak $H_0$ dan menyimpulkan variabel bebas secara simultan berpengaruh terhadap variabel terikat pada setiap metode yang digunakan.

Tabel 5. Uji Wald Model Regresi Logistik

| | Tanpa *resampling* | *Combine undersampling* | SMOTE tomke *link* |
|---|---|---|---|
| *Intercept* | -6,34 | 0,151 | 0,142 |
| $X_1$ | -1,518 | -1,913 | -3,245 |
| $X_2$ | 0,002 | 0,572 | 0,909 |
| $X_3$ | -3,605 | -3,74 | -9,304 |
| $X_4$ | -0,653 | -0,7 | -0,914 |
| $X_5$ | 0,557 | 0,008 | 1,548 |
| $X_6$ | 1,436 | 1,616 | 1,513 |

Dari Tabel 5 dapat dilihat pada model regresi logistik tanpa *resampling* dan *combined undersampling*, diperoleh bahwa hanya variabel bebas $X_3$ memiliki nilai $|W|$ > 1,959 yang berarti variabel bebas $X_3$ berpengaruh secara signifikan terhadap variabel terikat. Pada model regersi logistik SMOTE Tomek *link* diperoleh bahwa variabel bebas $X_1$ dan $X_3$ memiliki nilai $|W|$ > 1,959 yang berarti keduanya berpengaruh secara signifikan terhadap variabel terikat. Dari keterangan di atas diketahui bahwa tidak semua variabel X berpengaruh signifikan pada semua metode resampling maka dilakukan pemodelan ulang regresi logistik. Dari sini kemudian dilakukan pemodelan ulang dengan hanya menyertakan variabel signifikan.

Tanpa resampling:

$$P(Y=1)=\pi(x)=\frac{\exp(-1,636-1,172X_3)}{1+\exp(-1,636-1,172X_3)}$$

*Combine undersampling*:

$$P(Y=1)=\pi(x)=\frac{\exp(0,21-1,208X_3)}{1+\exp(0,21-1,208X_3)}$$

SMOTE tomek *link*:

$$P(Y=1)=\pi(x)=\frac{\exp(0,386-0,428X_1-1,553X_3)}{1+\exp(0,386-0,428X_1-1,553X_3)}$$

Setelah pembuatan model dengan regresi logistik dari data hasil klasifikasi tanpa *resampling*, *combined undersampling* dan SMOTE+tomek *link* dilanjutkan dengan membuat *confusion* matrix

Tabel 6. *Confusion* Matrix

| Metode | Hasil Observasi | Taksiran klasifikasi | |
|---|---|---|---|
| | | Mayoritas | Minoritas |
| Tanpa *Resampling* | Mayoritas | 90 | 0 |
| | Minoritas | 23 | 0 |
| *Combine Undersampling* | Mayoritas | 24 | 66 |
| | Minoritas | 5 | 18 |
| SMOTE Tomek *Link* | Mayoritas | 56 | 34 |
| | Minoritas | 14 | 9 |

Dari Tabel 6 dapat dihitung performa klasifikasi berupa nilai *Accuracy*, *Sensitivity*, *Specificity*, G-*mean* dan AUC seperti pada Tabel 7.

Tabel 7. Perbandingan Performa Klasifikasi

| Performa klasifikasi | Tanpa *resampling* | *Combined undersampling* | SMOTE *tomek link* |
|---|---|---|---|
| Akurasi | 85,840 | 37,168 | 57,522 |
| Sensitivitas | 0,00 | 78,261 | 39,130 |
| Spesifisitas | 100,00 | 26,667 | 62,222 |
| G-*Mean* | 0,00 | 45,683 | 49,343 |
| AUC | 50,00 | 52,464 | 50,676 |

Berdasarkan Tabel 4.10 dari segi akurasi, model yang diperoleh tanpa *resampling* mampu memprediksi lebih dari 85% dari data yang ada sedangkan model dari *combine undersampling* dan SMOTE tomek *link* masing-masing hanya mampu memprediksi setidaknya 37% dan 57% dari data. Dari segi sensitivitas, model yang diperoleh tanpa *resampling* sepenuhnya tidak dapat memprediksi data kelas minoritas sedangkan model dari *combine undersampling* dan SMOTE+tomek *link* masing-masing setidaknya mampu memprediksi 78% dan 39% data dari kelas minoritas.

Dari segi spesifisitas, model yang diperoleh tanpa *resampling* mampu memprediksi sepenuhnya data kelas mayoritas sedangkan model dari *combine undersampling* dan SMOTE+tomek *link* masing-masing hanya dapat memprediksi setidaknya 27% dan 62% dari data kelas mayoritas. Dari segi G-mean, model yang diperoleh tanpa *resampling* sepenuhnya tidak seimbang dalam memprediksi data kelas minoritas dan mayoritas sedangkan model *combine undersampling* dan SMOTE tomek *link* masing-masing seimbang setidaknya pada tingkat 46% dan 49% dalam memprediksi data baik data kelas minoritas maupun mayoritas. Dari segi *area under curve* (AUC), model tanpa *resampling* nilai AUC nya adalah 50,00 dimana lebih rendah dari model *combine undersampling* dan SMOTE+tomek *link* dengan masing-masing nilai AUC 52,46 dan 50,68.

Dari hasil performa klasifikasi di atas jika dilihat dari sensitivitas dan AUC maka hasil prediksi dengan *resampling combine undersampling* adalah yang terbaik karena lebih mampu memprediksi kelas minoritas dibandingkan dua teknik lainya.

$$P(Y=1) = \pi(x) = \frac{\exp(0{,}21-1{,}208X_3)}{1+\exp(0{,}21-1{,}208X_3)} = \frac{2{,}7182838^{(0{,}21-1{,}208X_3)}}{1+2{,}7182838^{(0{,}21-1{,}208X_3)}}$$

dari model combine undersampling di atas dapat diketahui bahwa peluang belajar minimal 3 jam untuk siswa yang suka belajar daring (X3=1) = 0,27 yang berarti peluang untuk siswa yang suka belajar daring untuk belajar kurang dari 3 jam =1-0,27=0,73. Peluang siswa yang tidak suka belajar daring untuk tetap belajar minimal 3 jam (X3=0) = 0,55 yang berarti peluang untuk siswa yang tidak suka belajar daring untuk belajar kurang dari 3 jam =1-0,55=0,45. Ketika menerapkan teknik resampling combine

undersampling pada data total tanpa pembagian maka diperoleh *confusion* matrix berikut

Tabel 8. *Confusion* Matrix *Combine Undersampling*

| Metode | Hasil observasi | Taksiran klasifikasi | |
|---|---|---|---|
| | | Mayoritas | Minoritas |
| Combine Undersampling | Mayoritas | 44 | 102 |
| | Minoritas | 18 | 128 |

Dari *confusion* matrix di Atas kemudian dapat dihitung performa klasifikasi pada data total dengan nilai performa klasifikasi dari penerapan teknik resampling *combine undersampling* sebagai berikut

Tabel 9. *Combine Undersampling* Data Total

| Performa Klasifikasi | Combine Undersampling |
|---|---|
| Akurasi | 58,90 |
| Sensitivitas | 87,67 |
| Spesifisitas | 30,14 |
| G-*Mean* | 54,40 |
| AUC | 58,90 |

## KESIMPULAN

Berdasarkan hasil pembahasan pada bab sebelumnya dapat diperoleh beberapa kesimpulan dari penelitian ini antara lain sebagai berikut: 1) Klasifikasi tanpa resampling memiliki nilai accuracy yang tinggi namun sepenuhnya mengabaikan kelas minoritas dilihat dari sensitivity dan G-mean nya. Hal ini sesuai dengan apa yang diungkapkan Verdikha dkk pada penelitian sebelumnya bahwasanya nilai accuracy pada klasifikasi pada data tidak seimbang menempatkan lebih banyak bobot pada kelas mayoritas. Nilai sensitivity dan G-mean menunjukan bahwa penerapan teknik resampling mampu meningkatkan akurasi prediksi terhadap kelas minoritas. Ditinjau dari nilai AUC yang dihasilkan teknik resampling *combine undersampling* adalah yang paling tinggi dengan nilai 52,46. 2) *combine undersampling* memberikan kinerja terbaik dalam meningkatkan akurasi prediksi pada kelas minoritas dengan variabel signifikan adalah $X_3$ (kesukaan siswa terhadap pembelajaran daring). Adapun model yang diperoleh adalah:

$$P(Y=1) = \pi(x) = \frac{\exp(0{,}21-1{,}208X_3)}{1+\exp(0{,}21-1{,}208X_3)} = \frac{2{,}7182838^{(0{,}21-1{,}208X_3)}}{1+2{,}7182838^{(0{,}21-1{,}208X_3)}}$$

dari model *combine undersampling* di atas dapat diketahui bahwa peluang belajar minimal 3 jam untuk siswa yang suka belajar daring ($X_3$=1) = 0,27 yang berarti peluang untuk siswa yang suka belajar daring untuk belajar kurang dari 3 jam =1-0,27=0,73. Peluang siswa yang tidak suka belajar daring untuk tetap belajar minimal 3 jam ($X_3$=0) = 0,55 yang berarti peluang untuk siswa yang tidak suka belajar daring untuk belajar kurang dari 3 jam =1-0,55=0,45.

## DAFTAR PUSTAKA

Apriana, S. D. E. 2016. *Rare Event Weighted Logistic Regression Untuk Klasifikasi Imbalanced Data* [Tesis]. Surabaya: Institut Teknologi Sepuluh Nopember, Fakultas Matematika Dan Ilmu Pengetahuan Alam.

Chawla, Bowyer, Hall, and Kegelmeyer. (2002) "SMOTE: Synthetic Minority Oversampling Technique". Journal of Artificial Intelligence Research 16. Page 321-357.

Faisal, M.R. & Nugrahadi, D. T. 2016. Belajar Data Science Klasifikasi Dengan Bahasa Pemrograman R. Banjarbaru: Scripta Cendekia.

Hart, P.E (1967) 'The condensed nearest Neighbor Rule',pp. 1966-1967.

Insani, F., AF, S., & LP, T. 2015. Metode Bootstrap Aggregating Regresi. Logistik untuk Peningkatan Ketepatan Klasifikasi Regresi Logistik Ordinal [abstrak]. Makassar: Unhas, Matematika dan Ilmu Pengetahuan Alam.

Misna, Rais dan Utami, I. T. 2018. Analisis Regresi Logistik Biner Untuk Mengklasifikasi Penderita Hipertensi Berdasarkan Kebiasaan Merokok Di RSU Mokopido Toli-Toli. Science and Technology, 7(3), 2.

Shintia, R.2018 Penerapan combine undresampling pada klasifikasi data imbalanced biner (studi kasus : desa tertinggal di Jawa timur tahun 2014) [skripsi] Surabaya : Institut Teknologi Sepuluh November, Fakultas Matematika, komputasi dan sains data.

Slameto.2003. *Belajar Dan Faktor Yang Mempengaruhinya*. Jakarta:Rineka Cipta.

Tomek, I. (1976).'Two Modification of CNN',IEEE Transaction on System, Man and Cybernetics, pp. 769-772

Yu, X., Zhou, M., Chen, X., Deng, L., & Wang, L. (2017). Using Class Imbalance Learning for Cross-Company Defect Prediction.