Jurnal TRANSFORMATIKA
Vol.21, No.2, Januari 2024, pp. 73 - 83
P-ISSN: 1693-3656, E-ISSN: 2460-6731
https://journals.usm.ac.id/index.php/transformatika/

Random State Parameter Undersampling untuk Penanganan Data dengan Kelas Tidak Seimbang pada Algoritme Random Forest

Galet Guntoro Setiaji*, Joko Suntoro[2], Ahmad Rifa'i[3]
[1] Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang
Jl. Soekarno Hatta, Semarang, telp:024-6702757, e-mail: gallet@usm.ac.id
[2]Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang
Jl. Soekarno Hatta, Semarang, telp:024-6702757, e-mail: jokosuntoro@usm.ac.id
[3]Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang
Jl. Soekarno Hatta, Semarang, telp:024-6702757, e-mail: rifai@usm.ac.id

ARTICLE INFO

History of the article :
Received 01 March 2024
Received in revised form 15 March 2024
Accepted 25 March 2024
Available online 27 March 2024

Keywords:

Classification; Class imbalanced; Random Forest; Undersampling; Random State

*** Correspondece:**

Telepon:
+62 81229219880

E-mail:
gallet@usm.ac.id

ABSTRACT (10 PT)

Algoritme Random Forest (RF) sangat populer digunakan pada metode klasifikasi karena waktu learning yang cepat, mampu melakukan pembobotan pada variabel, dan kinerja yang sangat baik pada dataset berukuruan besar, namun algoritme RF mempunyai performa yang buruk saat menangani data dengan kelas tidak seimbang. Data dengan kelas tidak seimbang adalah jumlah data pada kelas tertentu lebih banyak dibandingkan dengan jumlah data pada kelas lainnya. *Undersampling* (US-RF) adalah salah satu metode yang digunakan untuk penanganan data dengan kelas tidak seimbang, namun metode undersampling akan memilih dan mereduksi data secara acak pada kelas mayoritas sehingga berakibat hilangnya data yang berpotensi berguna. Untuk menghindari hilangnya data yang berpotensi berguna tersebut karena dipilih secara acak, maka akan diterapkan penetapan nilai random state pada metode undersampling. Metode yang diusulkan diberi nama random state parameter *undersampling Random Forest* (RSUS-RF). Dalam penelitian ini akan dibandingkan antara metode RF, US-RF dan RSUS-RF. Hasil penelitian menunjukkan nilai rata-rata akurasi metode RSUS-RF lebih tinggi dibandingkan dengan metode RF dan US-RF dengan nilai rata-rata akurasi metode RSUS-RF sebesar 0.8259, sedangkan nilai rata-rata akurasi metode RF sebesar 0.8035 dan metode US-RF sebesar 0.7945. Serta terdapat perbedaan secara signifikan diantara ketiga metode tersebut ketika diuji menggunakan Friedman Test dengan nilai p-value adalah 0.005.

1. INTRODUCTION

Metode klasifikasi banyak digunakan peneliti pada bidang data mining dan machine learning karena mudah diterapkan untuk banyak bidang [1]. Salah satu algoritme pada metode klasifikasi yang sering digunakan adalah *algoritme Random Forest* [2]. *Algoritme Random*

Forest digunakan dalam beberapa bidang penelitian seperti bidang lingkungan hidup [3], bidang kesehatan [4], [5], bidang teknik [6], bidang ilmu sosial [7], bidang *neuroscience* [8], dan bidang energi [9].

Algoritme Random Forest pertama kali diperkenalkan oleh *Leo Breiman* [10], *algoritme Random Forest* sangat populer digunakan pada metode klasifikasi karena waktu learning yang cepat, mampu melakukan pembobotan pada variabel, dan kinerja yang sangat baik pada dataset berukuruan besar [11] [12]. Namun *algoritme Random Forest* mempunyai performa yang buruk saat menangani data dengan kelas tidak seimbang [13][14]. Data dengan kelas tidak seimbang diartikan bahwa jumlah data pada kelas tertentu lebih banyak dibandingkan dengan jumlah data pada kelas lainnya [15]–[17]. Data pada kelas yang lebih sedikit disebut dengan kelas minoritas, sedangkan data dengan kelas yang lebih banyak disebut dengan kelas mayoritas [18].

Data dengan kelas tidak seimbang dapat diselesaikan dengan dua metode, yaitu metode internal (modifikasi algoritme) dan metode eksternal (pengolahan data awal) [19]. Metode internal yang pernah diusulkan oleh para peneliti adalah *One-Class SVM* [20], *SVM-Forest* [21], *Least Square-SVM* [22] dan *Fuzzy SVM* [23]. Sedangkan pada metode eksternal (pengolahan data awal) para peneliti menggunakan metode *oversampling* [24], *undersampling* [25][26] dan *SMOTE* [27][28].

Metode *oversampling* pernah diusulkan oleh peneliti [29]. Pada *oversampling*, data minoritas akan disalin beberapa kali secara acak, sehingga kedua kelas menjadi seimbang, para peneliti juga menggunakan *SMOTE* [28], *SMOTE* adalah pengembangan dan kombinasi dari undersampling dan oversampling [27]. Metode undersampling mempunyai kelebihan yaitu sangat mudah diterapan dan digunaan untuk menangani dataset besar yang mempunyai kelas tidak seimbang [30]. Hasil kinerja metode undersampling lebih baik daripada metode oversampling untuk menangani data dengan kelas tidak seimbang [31]. Metode *undersampling* tepat diterapkan pada algoritme klasifikasi [32] dan metode undersampling lebih stabil daripada metode *oversampling* dan SMOTE [33], namun metode undersampling akan memilih dan mereduksi data secara acak pada kelas mayoritas [34] sehingga berakibat hilangnya data yang berpotensi berguna [35].

Pada penelitian ini, data pada kelas mayoritas tidak dihilangkan/direduksi secara acak, karena pemilihan data secara acak dapat mengakibatkan hilangnya data yang berpotensi berguna [35]. Untuk menghindari hilangnya data yang berpotensi berguna tersebut, maka akan diterapkan penetapan nilai *random state* pada metode *undersampling*. Metode dalam penelitian ini diberi nama *random state parameter undersampling random forest*. Tujuan dalam penelitian ini adalah menerapkan sebuah metode *random state parameter undersampling* yang akan diterapkan pada *algoritme Random Forest* untuk menyelesaikan masalah data dengan kelas tidak seimbang.

RESEARCH METHODS

Menurut metode, penelitian ini menggunakan metode penelitian eksperimen. Menurut Dawson [36], metode penelitian eksperimen adalah ujicoba yang dikontrol oleh peneliti sendiri untuk melakukan investigasi hubungan kausal (hubungan sebab-akibat). Langkah-langkah penelitian dapat dilihat pada Figure 1, yang berisi sebagai berikut: (1) Tinjauan Pustaka, (2) Pengumpulan Dataset, (3) Metode yang Diusulkan, (4) Eksperimen dan Pengujian Metode, (5) Evaluasi hasil.

Analisis Permasalahan dan Tinjauan Pustaka

Algoritme Random Forest pertama kali diperkenalkan oleh Breiman [37]. *Random Forest* bersifat *ensemble*, dimana setiap pengklasifikasi pohon keputusan (*decision tree*) akan menciptakan sebuah hutan (*forest*). Dari seluruh pohon keputusan tersebut akan dipilih jumlah suara terbanyak (*voting*) yang akan digunakan untuk menentukan kelas dari sebuah data inputan [38]. Hal ini secara langsung dapat mengatasi masalah ketika melakukan klasifikasi hanya menggunakan satu pohon keputusan saja sering kali tidak optimal, tetapi dengan memasukkan banyak pohon keputusan, maka akan diperoleh nilai akurasi yang optimal secara global.
Algoritme Random Forest adalah gabungan dari *algoritme Bagging* dan metode *random vector* [39]. Setiap pohon dibangun dari sample bootstrap yang berasal dari dataset asli. *Random Forest* adalah pengklasifikasi berbentuk pohon {h(x,θ k),k=1,…} dimana θ k adalah *random vector* yang diditribusikan secara independen dan masing masing tree.

Metode *undersampling* 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. Langkah pertama pada metode *undersampling* adalah pemilihan dataset kemudian dihitung selisih antara kelas mayoritas dan minoritas, jika masih terdapat selisih antara jumlah kelas maka dataset kelas mayoritas akan dihapus secara acak sampai jumlah kelas mayoritas sama dengan kelas minoritas. Metode *undersampling* dapat lebih efektif dan cepat dalam proses pelatihan klasifikasi untuk data dengan kelas tidak seimbang. Figure 2 menunjukkan *flowchart* metode undersampling.

Pengumpulan *Dataset*

Dataset yang digunakan dalam penelitian ini adalah dataset publik bernama *Knowledge Extraction Evolutionary Learning* (KEEL) repository dan UCI *Machine Learning repository* yang digunakan oleh peneliti dengan topik data dengan kelas tidak seimbang. Dataset KEEL repository dapat diunduh melalui link https://sci2s.ugr.es/keel/imbalanced.php, sedangkan dataset UCI machine learning repository dapat diunduh melalui link https://archive.ics.uci.edu/ml/datasets.php. Table 1 menunjukkan dataset yang digunakan pada penelitian ini.

Metode Yang Diusulkan

Dataset yang mengandung data dengan kelas tidak seimbang dihitung jumlah data pada kelas minoritas dan jumlah data pada kelas mayoritas. Kemudian data pada kelas mayoritas akan direduksi secara acak menggunakan metode undersampling, agar jumlah data pada kelas mayoritas sama dengan jumlah data pada kelas minoritas. Untuk menghindari hasil akurasi yang berubah-ubah karena reduksi secara acak pada metode undersampling, maka ditetapkan nilai random state. Pengaturan nilai random state yang digunakan dalam penelitian ini adalah sama dengan jumlah data pada kelas minoritas. Setelah dataset tersebut seimbang, maka akan dilakukan klasifikasi menggunakan algoritme *Random Forest*. Metode yang diusulkan ini diberi nama *Random State Parameter Undersamping Random Forest* (RSUS-RF). Figure 3 menunjukkan *flowchart* metode yang diusulkan.

Eksperimen dan Pengujian Metode

Tahapan eksperimen pada penelitian ini adalah sebagai berikut:

a. Menyiapkan dan mengumpulkan dataset UCI *Machine Learning Repository* yang mengandung data dengan kelas tidak seimbang.

DOI : https://doi.org/ 10.26623/transformatika.v21i2.8901

b. Dataset UCI *Machine Learning Repository* dibagi menjadi dua bagian menggunakan metode 10-fold cross validation.
c. Melakukan pengujian data training dan data testing dengan menggunakan *algoritme Random Forest*, kemudian catat hasil akurasinya, metode ini disebut dengan RF.
d. Melakukan pengujian data training dan data testing dengan menggunakan algoritme Random State *Undersampling Random Forest*, kemudian catat hasil evaluasi, metode ini disebut dengan RSUS-RF.
e. Membandingkan hasil akurasi pada metode RF dan RSUS-RF dengan uji beda t-Test, kemudian mengambil hasil yang terbaik.

Dalam penelitian ini menggunakan program bantu Python versi 3.0, SPSS 16.0, XLSTAT 2016. Spesifikasi komputer yang digunakan dalam penelitian ini dapat dilihat pada Table 2.

Tabel 2. Spesifikasi Komputer yang Digunakan

Jenis	**Keterangan**
Processor	Intel Core i7-7500U CPU @ 2.70GHz (4 CPUs)
Memory	16 GB
Hardisk	SSD 1 TB
Sistem Operasi	Windows 10 Enterprise 64-bit
Aplikasi	Python versi 3.0, XLSTAT 2016

Pengukuran evaluasi pada penelitian ini menggunakan akurasi. Nilai akurasi didapatkan dari tabel confusion matrix. Confusion matrix adalah tabel yang berisi matriks 2 dimensi, salah satu dimensi menunjukkan nilai prediksi dari klasifikasi dan dimensi lainnya menunjukkan nilai aktual dari klasifikasi [40]. Table 3 menunjukkan confusion matrix dengan 2 kelas [41]. Jika nilai prediksi benar dan nilai aktual benar maka disebut *True Positive* (TP). Jika nilai prediksi salah dan nilai aktual salah maka disebut *True Negative* (TN). Jika nilai prediksi benar dan nilai aktual salah maka disebut *False Positive* (FP). Jika nilai prediksi salah dan nilai aktual benar maka disebut *False Negative* (FN).

Setelah nilai dari confusion matrix diketahui maka langkah selanjutnya adalah menghitung nilai evaluasi. Formula yang digunakan untuk menghitung nilai *evaluasi* [42] dapat dilihat pada persamaan 1.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

RESULTS

Dalam penelitian ini nilai akurasi dari metode yang diusulkan (RSUS-RF) akan dibandingkan dengan metode *Random Forest* (RF) dan metode *Undersampling Random Forest* (US-RF). Table 4 dan Figure 4 menunjukkan hasil pengukuran akurasi antar metode. Metode dengan cetak tebal diartikan bahwa metode tersebut lebih baik daripada metode lainnya.

Setelah didapatkan nilai akurasi masing-masing metode, maka langkah selanjutnya adalah dilakukan uji *Friedman Test*. Untuk mengetahui apakah terdapat perbedaan secara signifikan lebih dari dua metode maka digunakan uji *Friedman Test* [43]. Nilai α yang digunakan dalam penelitian ini adalah 0.05. Jika nilai p-value lebih kecil daripada nilai α, maka dapat disimpulkan bahwa terdapat perbedaan signifikan antar metode. Dan sebaliknya jika nilai p-value lebih besar daripada nilai α, maka dapat disimpulkan tidak terdapat perbedaan signifikan antar metode. Dalam penelitian ini didapatkan nilai *p-value* sebesar 0.005 sehingga dapat disimpulkan terdapat perbedaan secara

signifikan pada metode RF, US-RF, dan RSUS-RF. Table 5 menunjukkan hasil uji *Friedman Test* pada penelitian ini.

Dalam uji Friedman Test hanya mengetahui perbedaan signifikan semua metode, namun tidak diketahui urutan metode terbaik dan metode mana yang berbeda secara signifikan dengan metode lainnya, sehingga perlu dilakukan uji lanjutan yaitu uji *Nemenyi Post Hoc* [44]. Uji *Nemenyi Post Hoc* dihitung berdasarkan penghitungan *average rank* (AR) dan *critical difference* (CD) [45]. Table 6 menunjukkan hasil penghitungan AR, sedangkan nilai CD pada penelitian ini adalah 1.252. Nilai CD dihitung berdasarkan persamaan 1, dimana nilai q_q= 2.343, nilai k = 3 dan nilai N = 7. Dari penghitungan AR didapatkan hasil bahwa urutan metode terbaik adalah RSUS-RF, RF, kemudian US-RF.

$$CD = q_{\alpha}\sqrt{\frac{k(k+1)}{6N}} \quad (1)$$

Setelah diketahui nilai AR dan CD, maka langkah selanjutnya adalah dilakukan penghitungan pairwase comparison berdasarkan nilai *difference* (diff). Jika nilai diff lebih besar dari nilai CD, maka dapat disimpulkan bahwa terdapat perbedaan signifikan, begitu juga sebaliknya jika nilai diff lebih kecil dari nilai CD, maka dapat disimpulkan tidak terdapat perbedaan signifikan. Tabel 7 menunjukkan nilai diff penghitungan *Pairwase Comparison*, dan Tabel 8 menunjukkan nilai *significant difference*. Terlihat bahwa terdapat perbedaan signifikan antara metode RSUS-RF dengan metode RF dan metode US-RF. Sedangkan metode RF dan US-RF tidak terdapat perbedaan signifikan.

DISCUSSION

Hasil penelitian menunjukkan nilai rata-rata akurasi metode RSUS-RF lebih baik daripada metode RF dan US-RF dalam penanganan data dengan kelas tidak seimbang, dengan nilai rata-rata metode RSUS-RF sebesar 0.8259, sedangkan nilai rata-rata akurasi metode RF dan US-RF masing-masing sebesar 0.8035 dan 0.7945. Terdapat perbedaan signifikan antara ketiga metode ketika diuji dengan Friedman Test dengan nilai *p-value* sebesar 0.005. Penelitian ini telah memberikan kontribusi yaitu penanganan data dengan kelas tidak seimbang pada *algoritme Random Forest*. Data dengan kelas tidak seimbang dapat diselesaikan menggunakan metode *undersampling* dengan pengaturan nilai *random state*.

DOI : https://doi.org/ 10.26623/transformatika.v21i2.8901

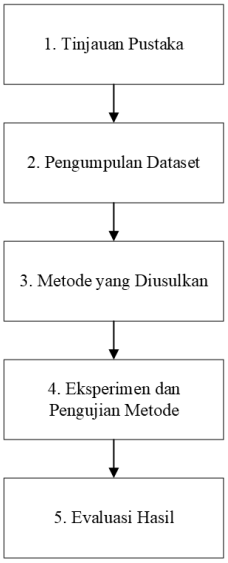

Figure 1. Tahapan Penelitian

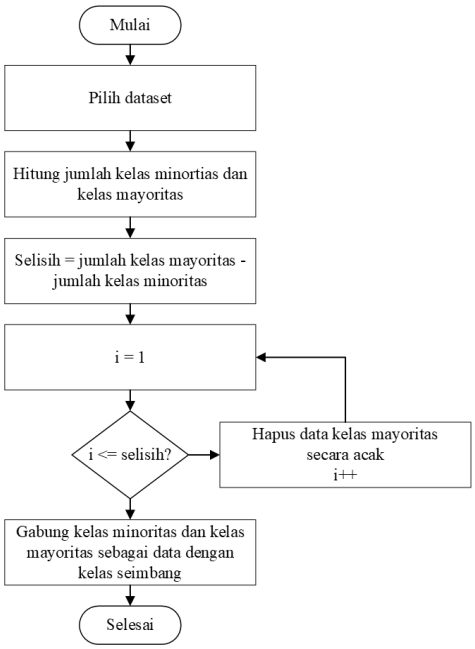

Figure 2. Show the flowchart of undersampling method

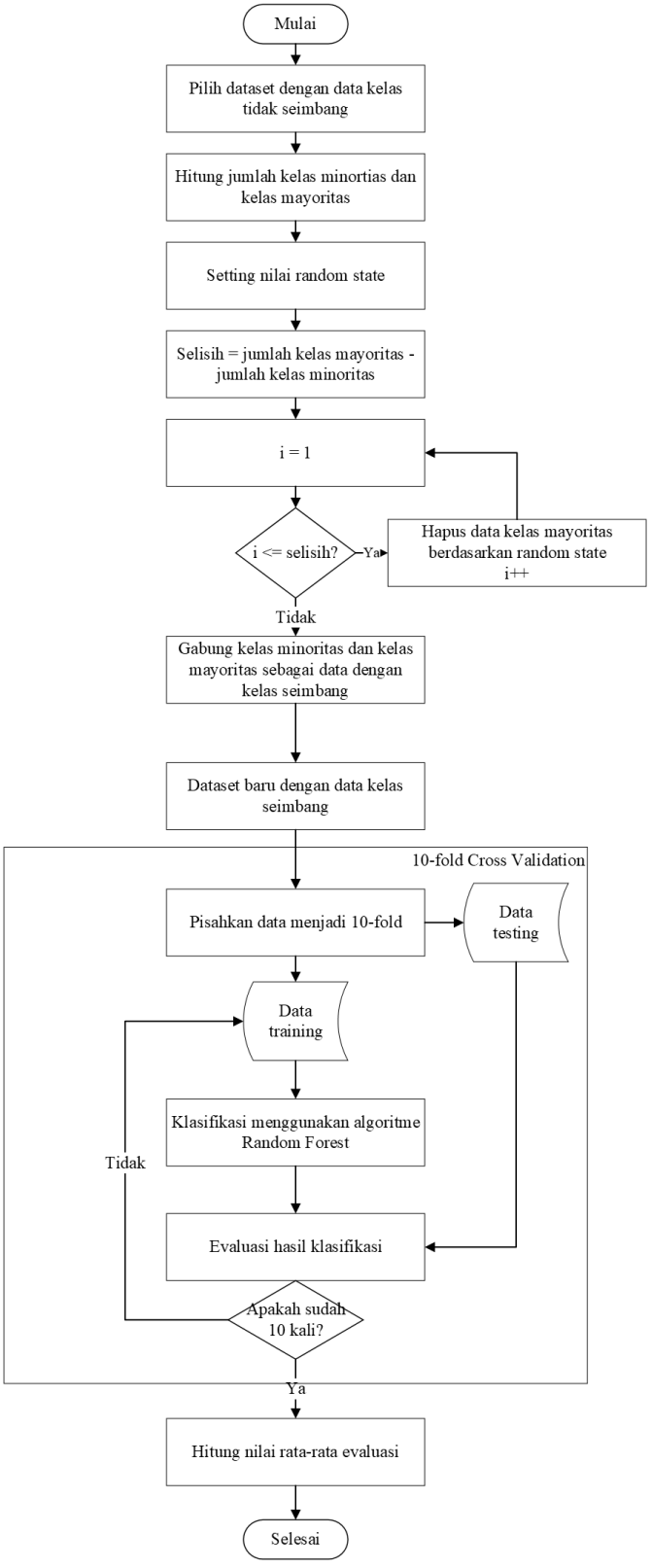

Figure 3. Flowchart Proposed Method

DOI : https://doi.org/ 10.26623/transformatika.v21i2.8901

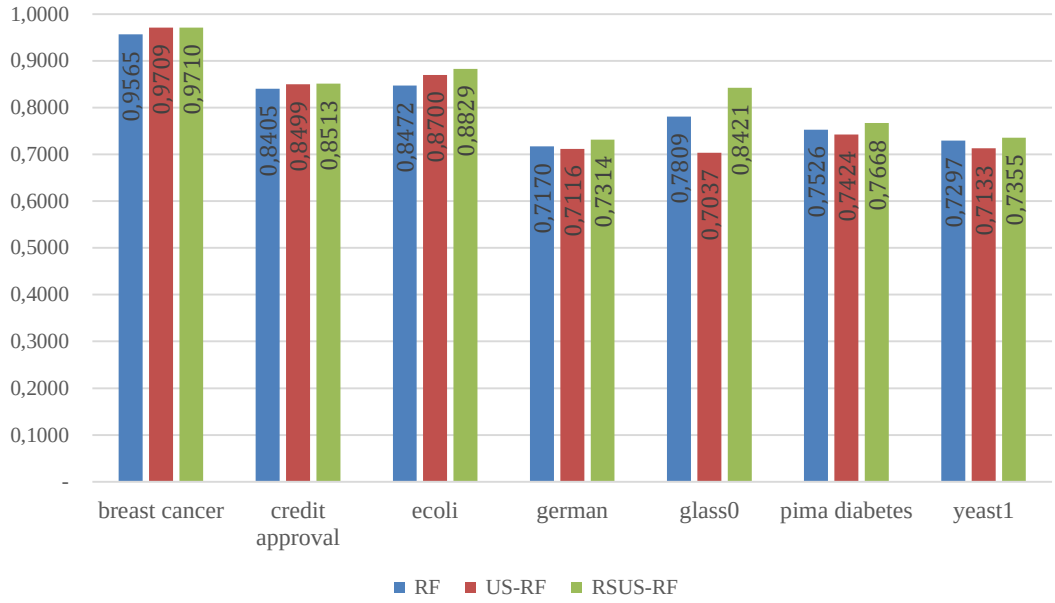

Figure 4. Diagram Nilai Akurasi Antar Metode

Table 1. Dataset dengan Kelas Tidak Seimbang

Repository	Dataset	Jumlah Data dalam Kelas Minoritas	Jumlah Data dalam Kelas Mayoritas	Imbalance Ratio
KEEL	ecoli1	77	259	3.36
KEEL	yeast1	429	1055	2.46
KEEL	pima	268	500	1.87
KEEL	glass1	76	138	1.82
UCI	german	300	700	2.33
UCI	breast cancer	241	458	1.90
UCI	credit approval	307	383	1.25

Table 2. Spesifikasi Komputer yang Digunakan

Processor	Intel® Core ™ i5-6200 CPU @ 2.30GHz 2.40GHz
Memory	16 GB
Hardisk	SSD 1 TB
Sistem Operasi	Windows 10 Enterprise 64-bit
Aplikasi	Python versi 3.0, SPSS 16.0, XLSTAT 2016

Table 3 Confusion Matrix

Keterangan	Prediksi Benar	Prediksi Salah
Aktual Benar	TP	FN
Aktual Salah	FP	TN

Table 4. Rekap Hasil Akurasi Antar Metode

Datasets	RF	US-RF	RSUS-RF (Proposed Method)

breast cancer	0.9565	0.9709	**0.9710**
credit approval	0.8405	0.8499	**0.8513**
ecoli1	0.8472	0.8700	**0.8829**
german	0.7170	0.7116	**0.7314**
glass0	0.7809	0.7037	**0.8421**
pima	0.7526	0.7424	**0.7668**
yeast1	0.7297	0.7133	**0.7355**

Table 5. Hasil Uji Friedman Antar Metode

CONCLUSIONS AND RECOMMENDATIONS

Penelitian ini telah memberikan kontribusi untuk menangani data dengan kelas tidak seimbang pada *algoritma Random Forest*. Namun ada beberapa metode yang dapat diuji coba lagi pada penelitian selanjutnya agar mendapatkan hasil yang lebih baik. Pada penelitian ini terdapat dataset yang mempunyai jumlah fitur yang banyak. Penelitian lanjutan mungkin dapat ditambahkan dengan metode seleksi fitur. Metode seleksi fitur dapat mengurangi dimensi data dan mengidentifikasi fitur yang relevan sehingga dapat membangun model pembelajaran yang kuat. Metode seleksi fitur yang pernah diusulkan oleh *Khoshgoftaar et al* adalah *chi-square, information gain, gain ratio, two types of ReliefF* (RF and RFW), dan *symmetrical uncertainty*.

REFERENCES

[1] M. L. Wong, K. Seng, and P. K. Wong, "Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain," *Expert Syst Appl*, vol. 141, Mar. 2020, doi: 10.1016/j.eswa.2019.112918.

[2] R. M. Mohana, C. K. K. Reddy, P. R. Anisha, and B. V. R. Murthy, "Random forest algorithms for the classification of tree-based ensemble," *Mater Today Proc*, Feb. 2021, doi: 10.1016/j.matpr.2021.01.788.

[3] F. Zhang and X. Yang, "Improving land cover classification in an urbanized coastal area by random forests: The role of variable selection," *Remote Sens Environ*, vol. 251, Dec. 2020, doi: 10.1016/j.rse.2020.112105.

[4] S. Asadi, S. E. Roshan, and M. W. Kattan, "Random forest swarm optimization-based for heart diseases diagnosis," *J Biomed Inform*, vol. 115, Mar. 2021, doi: 10.1016/j.jbi.2021.103690.

[5] B. O. Macaulay, B. S. Aribisala, S. A. Akande, B. A. Akinnuwesi, and O. A. Olabanjo, "Breast cancer risk prediction in African women using Random Forest Classifier," *Cancer Treat Res Commun*, vol. 28, Jan. 2021, doi: 10.1016/j.ctarc.2021.100396.

[6] M. Bassier, B. van Genechten, and M. Vergauwen, "Classification of sensor independent point cloud data of building objects using random forests," *Journal of Building Engineering*, vol. 21, pp. 468–477, Jan. 2019, doi: 10.1016/j.jobe.2018.04.027.

[7] J. Cho and S. Kim, "Personal and social predictors of use and non-use of fitness/diet app: Application of Random Forest algorithm," *Telematics and Informatics*, vol. 55, Dec. 2020, doi: 10.1016/j.tele.2019.101301.

[8] E. Feczko *et al.*, "Subtyping cognitive profiles in Autism Spectrum Disorder using a Functional Random Forest algorithm," *Neuroimage*, vol. 172, pp. 674–688, May 2018, doi: 10.1016/j.neuroimage.2017.12.044.

[9] R. H. Lin, Z. X. Pei, Z. Z. Ye, C. C. Guo, and B. D. Wu, "Hydrogen fuel cell diagnostics using random forest and enhanced feature selection," *Int J Hydrogen Energy*, vol. 45, no. 17, pp. 10523–10535, Mar. 2020, doi: 10.1016/j.ijhydene.2019.10.127.

[10] Y. M. Abd Algani, M. Ritonga, B. Kiran Bala, M. S. al Ansari, M. Badr, and A. I. Taloba, "Machine learning in health condition check-up: An approach using Breiman's random forest algorithm," *Measurement: Sensors*, vol. 23, Oct. 2022, doi: 10.1016/j.measen.2022.100406.

[11] R. O'Brien and H. Ishwaran, "A random forests quantile classifier for class imbalanced data," *Pattern Recognit*, vol. 90, pp. 232–249, Aug. 2019, doi: 10.1016/j.patcog.2019.01.036.

DOI : https://doi.org/ 10.26623/transformatika.v21i2.8901

[12] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.
[13] G. H. Fu, Y. J. Wu, M. J. Zong, and L. Z. Yi, "Feature selection and classification by minimizing overlap degree for class-imbalanced data in metabolomics," *Chemometrics and Intelligent Laboratory Systems*, vol. 196, Jan. 2020, doi: 10.1016/j.chemolab.2019.103906.
[14] R. O'Brien and H. Ishwaran, "A random forests quantile classifier for class imbalanced data," *Pattern Recognit*, vol. 90, pp. 232–249, Jun. 2019, doi: 10.1016/j.patcog.2019.01.036.
[15] B. S. Raghuwanshi and S. Shukla, "SMOTE based class-specific extreme learning machine for imbalanced learning," *Knowl Based Syst*, vol. 187, Jan. 2020, doi: 10.1016/j.knosys.2019.06.022.
[16] W. Lu, Z. Li, and J. Chu, "Adaptive Ensemble Undersampling-Boost: A novel learning framework for imbalanced data," *Journal of Systems and Software*, vol. 132, pp. 272–282, Oct. 2017, doi: 10.1016/j.jss.2017.07.006.
[17] H. Zhou, X. Dong, S. Xia, and G. Wang, "Weighted oversampling algorithms for imbalanced problems and application in prediction of streamflow[Formula presented]," *Knowl Based Syst*, vol. 229, Oct. 2021, doi: 10.1016/j.knosys.2021.107306.
[18] Z. Seng, S. A. Kareem, and K. D. Varathan, "A Neighborhood Undersampling Stacked Ensemble (NUS-SE) in imbalanced classification," *Expert Syst Appl*, vol. 168, Apr. 2021, doi: 10.1016/j.eswa.2020.114246.
[19] Y. M. Haibo He, "Imbalanced Learning," *The Institute of Electrical and Electronics Engineers, Inc. Published*, 2013, doi: 10.1002/9781118646106.ch5.
[20] S. Maldonado and J. López, "Imbalanced data classification using second-order cone programming support vector machines," *Pattern Recognit*, vol. 47, no. 5, pp. 2070–2079, 2014, doi: 10.1016/j.patcog.2013.11.021.
[21] G. Chen and Z. Ge, "SVM-tree and SVM-forest algorithms for imbalanced fault classification in industrial processes," *IFAC Journal of Systems and Control*, vol. 8, p. 100052, Jun. 2019, doi: 10.1016/j.ifacsc.2019.100052.
[22] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang, "New imbalanced fault diagnosis framework based on Cluster-MWMOTE and MFO-optimized LS-SVM using limited and complex bearing data," *Eng Appl Artif Intell*, vol. 96, Nov. 2020, doi: 10.1016/j.engappai.2020.103966.
[23] J. Liu, "Fuzzy support vector machine for imbalanced data with borderline noise," *Fuzzy Sets Syst*, vol. 413, pp. 64–73, Jun. 2021, doi: 10.1016/j.fss.2020.07.018.
[24] D. Lee and K. Kim, "An efficient method to determine sample size in oversampling based on classification complexity for imbalanced data," *Expert Syst Appl*, vol. 184, Dec. 2021, doi: 10.1016/j.eswa.2021.115442.
[25] W. Lu, Z. Li, and J. Chu, "Adaptive Ensemble Undersampling-Boost: A novel learning framework for imbalanced data," *Journal of Systems and Software*, vol. 132, pp. 272–282, Aug. 2017, doi: 10.1016/j.jss.2017.07.006.
[26] J. Hoyos-Osorio, A. Alvarez-Meza, G. Daza-Santacoloma, A. Orozco-Gutierrez, and G. Castellanos-Dominguez, "Relevant information undersampling to support imbalanced data classification," *Neurocomputing*, vol. 436, pp. 136–146, Aug. 2021, doi: 10.1016/j.neucom.2021.01.033.
[27] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution," *Inf Sci (N Y)*, vol. 512, pp. 1214–1233, Feb. 2020, doi: 10.1016/j.ins.2019.10.048.
[28] M.-J. Kim, D.-K. Kang, and H. B. Kim, "Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction," *Expert Syst Appl*, vol. 42, no. 3, pp. 1074–1082, 2015, doi: 10.1016/j.eswa.2014.08.025.
[29] M. a H. Farquad and I. Bose, "Preprocessing unbalanced data using support vector machine," *Decis Support Syst*, vol. 53, no. 1, pp. 226–233, 2012, doi: 10.1016/j.dss.2012.01.016.
[30] M. Koziarski, "Radial-Based Undersampling for imbalanced data classification," *Pattern Recognit*, vol. 102, Jun. 2020, doi: 10.1016/j.patcog.2020.107262.
[31] X. Xie, H. Liu, S. Zeng, L. Lin, and W. Li, "A novel progressively undersampling method based on the density peaks sequence for imbalanced data," *Knowl Based Syst*, vol. 213, Aug. 2021, doi: 10.1016/j.knosys.2020.106689.
[32] G. Kim, B. K. Chae, and D. L. Olson, "A support vector machine (SVM) approach to imbalanced datasets of customer responses: Comparison with other customer response models," *Service Business*, vol. 7, no. 1, pp. 167–182, 2013, doi: 10.1007/s11628-012-0147-9.

[33] M. Bach, A. Werner, and M. Palt, "The proposal of undersampling method for learning from imbalanced datasets," in *Procedia Computer Science*, 2019, vol. 159, pp. 125–134. doi: 10.1016/j.procs.2019.09.167.
[34] I. H. Laradji, M. Alshayeb, and L. Ghouti, "Software defect prediction using ensemble learning on selected features," *Inf Softw Technol*, vol. 58, pp. 388–402, 2015, doi: 10.1016/j.infsof.2014.07.005.
[35] Y. Qian, Y. Liang, M. Li, G. Feng, and X. Shi, "A resampling ensemble algorithm for classification of imbalance problems," *Neurocomputing*, vol. 143, pp. 57–67, 2014, doi: 10.1016/j.neucom.2014.06.021.
[36] C. W. Dawson, *Projects in Computing and Information Systems*, vol. 2. 2011.
[37] L. Breiman, "Random Forests," 2001.
[38] J. Han, M. Kamber, and J. Pei, "Data Mining Concept and Techniques Third Edition," 2012.
[39] C. Sammut and G. I. Webb, *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.
[40] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Inf Sci (N Y)*, vol. 340–341, pp. 250–261, 2016, doi: 10.1016/j.ins.2016.01.033.
[41] E. Alpaydın, *Introduction to machine learning*, 2nd ed., vol. 1107. London, England: The MIT Press Cambridge, Massachussetts, 2014. doi: 10.1007/978-1-62703-748-8-7.
[42] Y. Liu, X. Yu, J. X. Huang, and A. An, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," *Inf Process Manag*, vol. 47, no. 4, pp. 617–631, 2011, doi: 10.1016/j.ipm.2010.11.007.
[43] J. Suntoro, A. Ilham, and H. A. D. Rani, "New Method Based Pre-Processing to Tackle Missing and High Dimensional Data of CRISP-DM Approach," in *Journal of Physics: Conference Series*, 2020, vol. 1471, no. 1, p. 12012.
[44] M. Zheng *et al.*, "UFFDFR: Undersampling framework with denoising, fuzzy c-means clustering, and representative sample selection for imbalanced data classification," *Inf Sci (N Y)*, vol. 576, pp. 658–680, Oct. 2021, doi: 10.1016/j.ins.2021.07.053.
[45] N. Hidayati, J. Suntoro, and G. G. Setiaji, "Perbandingan Algoritma Klasifikasi untuk Prediksi Cacat Software dengan Pendekatan CRISP-DM," *Jurnal Sains dan Informatika*, vol. 7, no. 2, pp. 117–126, Nov. 2021, doi: 10.34128/jsi.v7i2.313.