JIKO (Jurnal Informatika dan Komputer)
Vol. 4, No. 1, April 2021, hlm. 38-43
DOI: 10.33387/jiko

Akreditasi KEMENRISTEKDIKTI, No.36/E/KPT/2019
p-ISSN: 2614-8897
e-ISSN: 2656-1948

PENERAPAN TEKNIK KOMBINASI OVERSAMPLING DAN UNDERSAMPLING UNTUK MENGATASI PERMASALAHAN *IMBALANCED DATASET*

Ariani Indrawati

Lembaga Ilmu Pengetahuan Indonesia
email: indrawati.ariani@gmail.com

(Naskah masuk: 12 Desember 2020, diterima untuk diterbitkan: 4 Februari 2021)

Abstrak

Salah satu permasalahan pada *machine learning* yang cukup sering terjadi adalah ketidakseimbangan data yang digunakan atau sering disebut dengan *imbalanced dataset*. Cukup banyak penelitian yang melaporkan bahwa *imbalanced dataset* ini seringkali memberikan hasil yang keliru. Perlu ada penanganan khusus sebelum *imbalanced dataset* tersebut dapat digunakan pada *machine learning*. Cara paling populer dan efektif dalam mengatasi permasalahan *imbalanced dataset* adalah melakukan *resampling*, baik *oversampling*, *undersampling*, ataupun kombinasi keduanya. Pada penelitian ini akan dilakukan uji coba teknik kombinasi dengan menggabungkan teknik *oversampling Synthetic Minority Oversampling Technique* (SMOTE) dengan teknik *undersampling Edited Nearest Neighbors* (ENN) dan *TomekLinks* terhadap *Support Vector Machine* (SVM). Tiga *public dataset* UCI yaitu *Breast Cancer Wisconsin*, *Pima Indian Diabetes*, dan *Heart Disease Detection* digunakan pada penelitian ini dengan *Python* sebagai alat bantu pemrograman. Berdasarkan hasil uji coba yang dilakukan diketahui bahwa teknik kombinasi dapat membantu mengatasi permasalahan *imbalanced dataset* pada *machine learning*, SMOTE-ENN dapat meningkatkan performa akurasi dari SVM sebesar 2% hingga 23%.

Kata kunci: *Imbalanced dataset, Resampling, Oversampling, Undersampling*

HYBRID OVERSAMPLING AND UNDERSAMPLING TECHNIQUES TO HANDLING IMBALANCED DATASET

Abstract

One of the problems in machine learning that often occurs is the imbalance of the data used or often called the imbalanced dataset. Quite a number of studies have reported that imbalanced dataset often gives false results. There needs to be special handling before the imbalanced dataset can be used in machine learning. The most popular and effective way to solve imbalanced dataset problems is resampling, either oversampling, undersampling, or a combination both of them. In this study, we tried a combination technique by combining Synthetic Minority Oversampling Technique (SMOTE) with the Edited Nearest Neighbors (ENN) and TomekLinks undersampling technique against the Support Vector Machine (SVM). We used three public UCI datasets, are Breast Cancer Wisconsin, Pima Indian Diabetes, and Heart Disease Detection in this study, and we also use Python as a programming tools. Based on the results of the experiment, it is known that the combination technique can solve the problem of imbalanced dataset in machine learning, SMOTE-ENN can improve the accuracy performance of the SVM by 2% to 23%.

Keywords: *Imbalanced dataset, Resampling, Oversampling, Undersampling*

1. PENDAHULUAN

Permasalahan *imbalanced dataset* seringkali kita jumpai pada berbagai domain, dimana jumlah data pada tiap kelas tidak seimbang. *Imbalanced* data terjadi ketika jumlah data dalam satu kelas jauh lebih tinggi (*majority class*) atau lebih rendah (*minority class*) dibandingkan kelas lainnya. *Imbalanced data* akan lebih sulit untuk dilakukan pengolahan dan analisis data seperti klasifikasi, pengklusteran, prediksi, dan sebagainya. Hal tersebut terjadi karena model analisis data tidak dirancang untuk mempertimbangkan distribusi kelas dalam meningkatkan akurasi dari model. Banyak penelitian yang melaporkan analisis data dengan *imbalanced data* seringkali memberikan hasil yang keliru.

Perlu ada penanganan khusus sebelum *imbalanced data* tersebut dapat digunakan untuk proses analisis data. Cara paling populer dalam mengatasi permasalahan *imbalanced* data adalah melakukan *resampling* dengan mengubah jumlah data pada tiap

kelas hingga mencapai jumlah data yang seimbang pada seluruh kelas dan merupakan sebuah teknik yang efektif [1, 2].

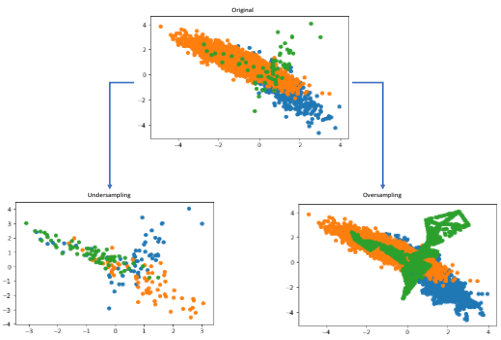
Banyak metode *resampling* telah diajukan baik *oversampling* maupun *undersampling*. Metode-metode untuk teknik *oversampling* antara lain, *Random Over Sampling* (ROS), *Synthetic Minority Oversampling Technique* (SMOTE) [3], *Borderline SMOTE* [4], *kMeans SMOTE* [5], *Support Vector Machine SMOTE* (SVM-SMOTE) [6], *Adaptive Synthetic* (ADASYN) [7], dll. Sedangkan untuk *undersampling* antara lain, *Random Under Sampling* (RUS), *TomekLinks* [8], *Edited Nearest Neighbors* (ENN) [9], One Side Selection (OSS) [10], Neighborhood Cleaning Rule (NCR) [11], dan sebagainya. Metode-metode *oversampling* atau *undersampling* tersebut telah diterapkan pada beberapa penelitian [12 – 18]

Namun, teknik *oversampling* dan *undersampling* memiliki celah atau kekurangan, *oversampling* yang dilakukan secara acak dapat mengakibatkan *overfitting* pada model yang dibentuk [1, 19], sedangkan *undersampling* dapat menghilangkan bagian-bagian penting pada *majority class* sehingga batas keputusan antar kelas lebih sulit dipelajari dan berpengaruh terhadap performa klasifikasi [7].

Dalam mengatasi kekurangan tersebut, pada penelitian ini akan dilakukan uji coba teknik kombinasi dengan menggabungkan teknik *oversampling* SMOTE dengan teknik *undersampling* ENN (SMOTE-ENN) dan TomekLinks (SMOTE-Tomek) untuk melakukan *resampling* terhadap 3 dataset UCI yaitu Breast Cancer Wisconsin, Pima Indian Diabetes, dan Heart Disease Detection. Selanjutnya dataset hasil *resampling* akan dilakukan klasifikasi menggunakan SVM. SVM dipilih karena beberapa penelitian menyatakan bahwa SVM merupakan metode paling baik dibandingkan dengan metode lainnya [20 - 25].

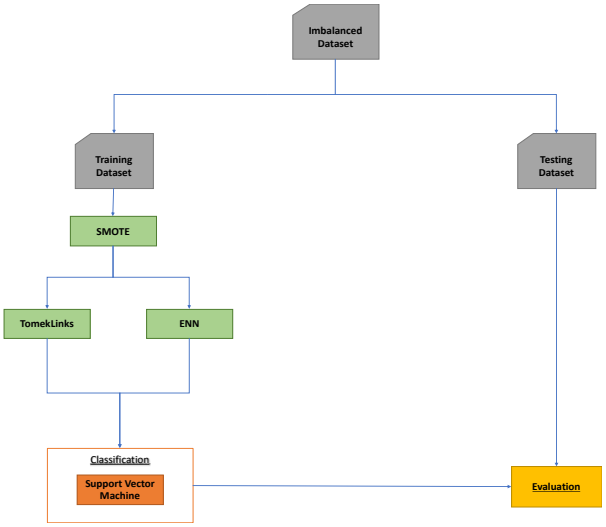
2. METODE PENELITIAN

Dalam melakukan *resampling* data terdapat 2 teknik, yaitu *oversampling* dengan menambahkan sejumlah data pada *minority class*, serta *undersampling* dengan mengurangi jumlah data pada *majority class*. Ilustrasi *oversampling* dan *undersampling* dapat dilihat pada gambar 1.



Gambar 1. Ilustrasi *oversampling* dan *undersampling*

Pada penelitian ini, langkap pertama adalah tahap pengumpulan *imbalanced data* yang kemudian akan dibagi 70% untuk pemodelan atau pelatihan dan 30% untuk data pengujian. Kemudian, data pelatihan akan dilakukan dilakukan *oversampling* data menggunakan metode SMOTE dan *undersampling* menggunakan TomekLinks dan ENN. Data hasil *resampling* tersebut dilakukan pemodelan klasifikasi menggunakan SVM. Tahap terakhir adalah pengujian *precision*, *recall*, dan *f-measure* dari pemodelan yang sudah terbentuk menggunakan data uji. Seluruh tahapan tersebut diproses dengan memanfaat Python sebagai alat bantu. *Source code* pada penelitian ini dapat diakses pada rin.lipi.go.id setelah artikel ini terbit. Gambar 2 mengilustrasikan tahapan yang akan dilakukan pada penelitian ini.

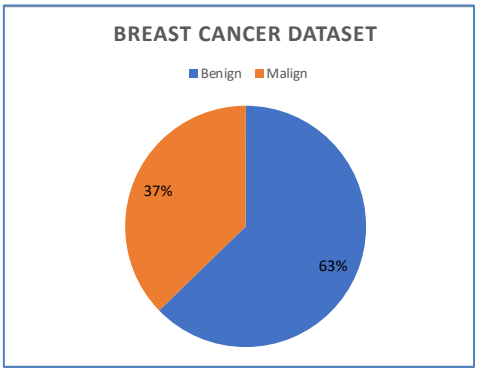


Gambar 2. Metode Penelitian

A. Dataset

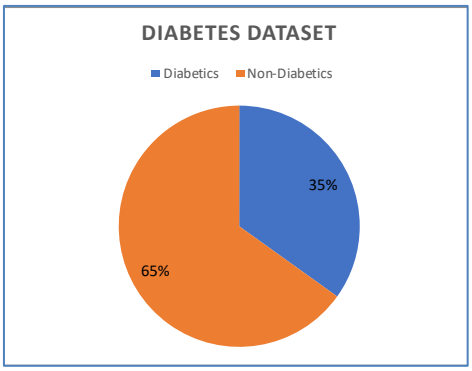
Penelitian ini menggunakan 3 dataset yang bersumber dari *UCI Machine Learning Repository* (https://archive.ics.uci.edu/ml/datasets.php).

a. *Breast cancer* dataset berjumlah 469 data yang dibagi menjadi 2 kelas, *Benign* dengan 357 data dan *Malign* dengan 212 data. presentase jumlah data dari *minority class* dan *majority class* dapat dilihat gambar 3.



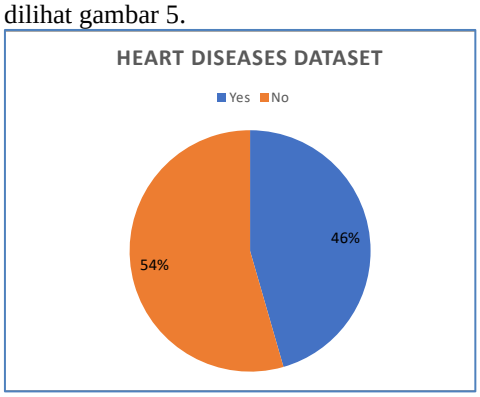
Gambar 3. Distribusi dataset *breast cancer*

b. *Pima Indian Diabetes* dataset berjumlah 768 data yang dibagi menjadi 2 kelas, *Diabetic* dengan 500 data dan *Non-diabetic* dengan 268 data. presentase jumlah data dari *minority class* dan *majority class* dapat dilihat gambar 4.



Gambar 4. Distribusi dataset *diabetes*

c. *Breast cancer* dataset berjumlah 303 data yang dibagi menjadi 2 kelas, pasien yang menderita penyakit jantung (kelas 1 / *Yes*) dengan 138 data dan pasien yang tidak menderita penyakit jantung (Kelas 0 / *No*) dengan 165 data. presentase jumlah data dari *minority class* dan *majority class* dapat dilihat gambar 5.



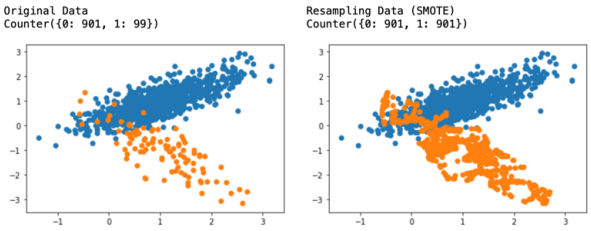
Gambar 5. Distribusi dataset *heart disesase*

B. Metode *Resampling*

Pada penelitian ini akan dilakukan ujicoba pengaruh teknik kombinasi SMOTE-ENN dan SMOTE-Tomek dimana *imbalanced data* akan dilakukan *oversampling* akan dilakukan dengan menggunakan SMOTE, kemudian data akan coba dikurangi menggunakan metode *undersampling* ENN dan TomekLinks. Bagian ini akan menyajikan penjelasan singkat mengenai metode SMOTE, ENN, dan TomekLink yang akan digunakan sebagai kombinasi *resampling* pada penelitian ini.

a. SMOTE

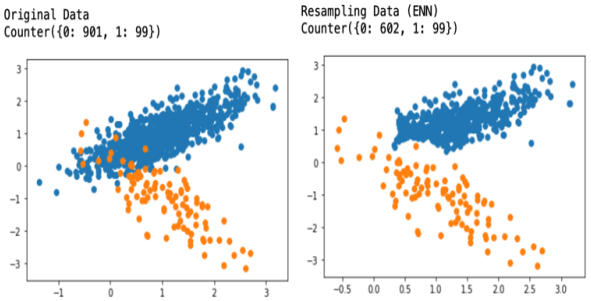
SMOTE dikenalkan pertama kali oleh Chawla pada tahun 2002. Cara kerja SMOTE adalah dengan menambahkan data-data buatan pada kelas *minority* dengan melakukan interpolasi pada data-data asli, sehingga data buatan yang dihasilkan sangat bervariasi. Algoritma dari SMOTE dapat dijelaskan [3]. Ilustrasi hasil *oversampling* dapat dilihat pada gambar 6, pada original data kelas 0 memiliki 901 data dan kelas 1 hanya memiliki 99 data, setelah dilakukan *oversampling* menggunakan SMOTE data pada kelas 1 bertambah menjadi 901 data.



Gambar 6. Ilustrasi SMOTE

b. ENN

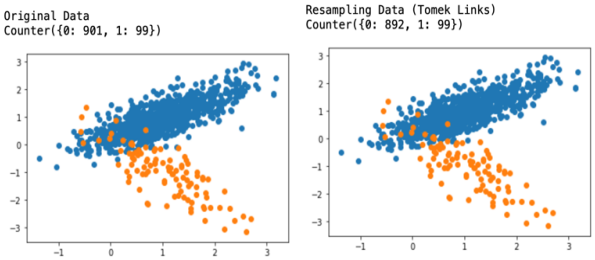
Dalam metode ENN yang diajukan oleh Wilson [9], *undersampling* pada *majority class* dilakukan dengan cara menghilangkan sampel dari *majority class* yang memiliki label berbeda pada data-data yang berdekatan. Metode ini sangat bergantung pada nilai *k* yang dipilih. Hasil *undersampling* diilustrasikan pada gambar 7, dapat dilihat bahwa data *majority class* (kelas 1) yang semula berjumlah 901 data berkurang menjadi 602 data.



Gambar 7. Ilustrasi ENN

c. TomekLink

TomekLinks diajukan oleh Tomek pada tahun 1972. Metode ini mengurangi sampel dengan cara 2 data yang saling berdekatan antara *minority class* dan *majority class*. Pada gambar 8, original data dan hasil resampling menggunakan TomekLink memang terlihat hampir sama, namun jika dihitung dari jumlah data *majority class* berkurang menjadi 892 data dari 901 data.



Gambar 8. Ilustrasi TomekLink

3. HASIL DAN PEMBAHASAN

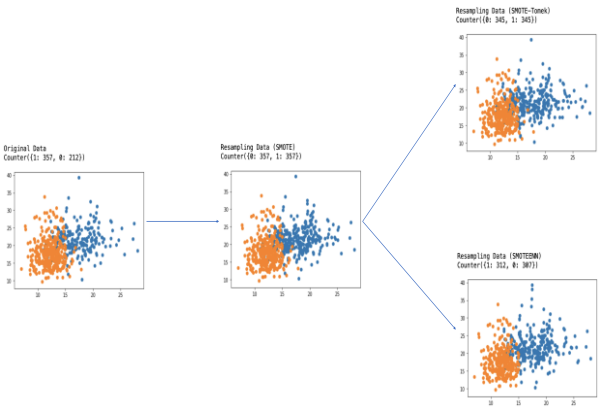
Pada bagian ini akan dijelaskan hasil eksperimen yang dilakukan pada penelitian ini untuk mengetahui pengaruh penerapan 2 metode kombinasi *oversampling* dan *undersampling*, yaitu SMOTE-ENN dan SMOTE-Tomek, terhadap performa metode klasifikasi SVM.

Langkah pertama yang dilakukan adalah melakukan *oversampling* pada setiap dataset menggunakan SMOTE. Data *breast cancer* terdapat penambahan data sebanyak 145 pada kelas *Malign*. Pada data diabetes mengalami penambahan data yang cukup siginifikan sebesar 232 pada kelas *Diabetic*, sedangkan pada data *heart disease* bertambah sebanyak 27 data kelas 1.

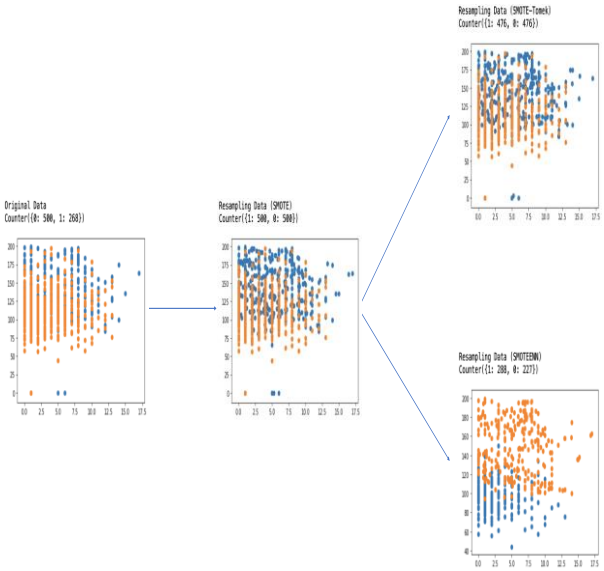
Langkah selanjutnya adalah melakukan proses *undersampling* dengan mengurangi hasil *oversampling* dengan metode *TomekLink* dan ENN pada kedua kelas baik *majority class* maupun *minority class*. Sebaran data dari proses *resampling* dapat dilihat pada gambar 9 untuk data *breast cancer*, gambar 10 untuk data *pima Indian diabetes*, dan gambar 11 untuk data *heart diseases*. Distribusi jumlah data pada masing-masing sebelum dan setelah proses *resampling* dengan teknik kombinasi dapat dilihat pada tabel 1.

Tabel 1. Jumlah Data Sebelum dan Sesudah *Resampling*

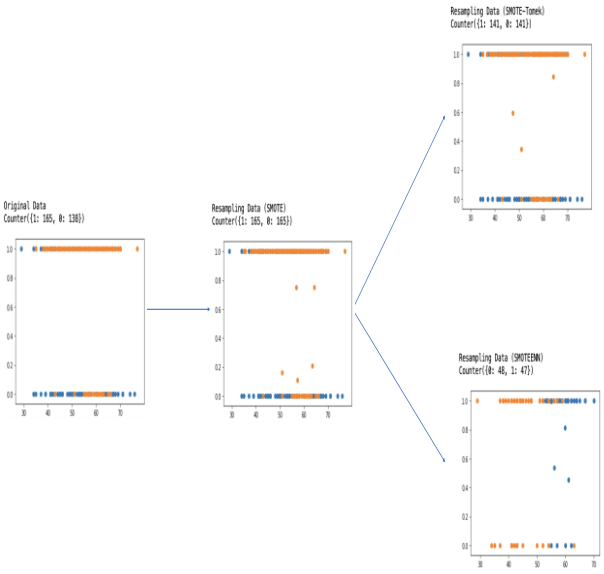
Dataset	Kelas	Jumlah Data			
		Original	*SMOTE*	*SMOTE-ENN*	*SMOTE-Tomek*
Breast Cancer	*Benign*	357	357	318	348
	Malign	212	357	303	348
Diabetes	*Diabetics*	268	500	296	471
	Non-Diabetics	500	500	225	471
Heart Disease	*Yes*	138	165	49	141
	No	165	165	44	141



Gambar 9. Sebaran data *breast cancer* tiap tahap *resampling*

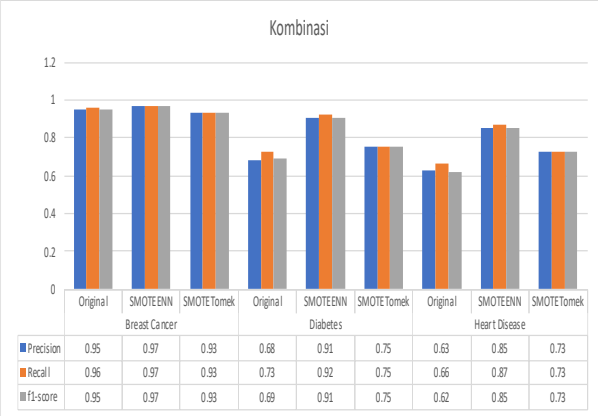


Gambar 10. Sebaran data *pima Indian diabetes* tiap tahap *resampling*



Gambar 11. Sebaran data *heart disease* tiap tahap *resampling*

Setelah dilakukan proses *resampling*, langkah selanjutnya adalah membagi data menjadi data latih dan data uji dengan perbandingan 70:30. Data latih akan digunakan untuk pemodelan SVM, sedangkan data uji akan digunakan untuk menguji *precision*, *recall*, dan *f-Measure* dari model yang dibentuk. Hasil ketiga nilai tersebut dapat dilihat pada gambar 12. Pada *breast cancer* dataset, *resampling* menggunakan SMOTE-ENN mampu menaikan *f-Measure* sebesar 0.02, namun sebaliknya SMOTE-Tomek menurunkan *f-Measure* menjadi dari 0.95 menjadi 0.93. Pada dataset *Pima Indian Diabetes*, kedua metode *resampling* dapat meningkatkan *f-Measure* sebesar 0.22 untuk SMOTE-ENN dan 0.06 untuk SMOTE-Tomek. Pada *heart disesases* dataset, SMOTE-ENN dan SMOTE-Tomek berhasil meningkatkan *f-Measure* sebesar 0.23 dan 0.11.



Gambar 12. *Precision*, *Recall*, dan *f-Measure* SMOTE-ENN dan SMOTE-Tomek

4. KESIMPULAN

Teknik *resampling* merupakan cara sederhana yang dapat membantu menangani permasalahan *imbalanced dataset* pada *machine learning*, baik *oversampling*, *undersampling*, maupun kombinasi keduanya. Hal tersebut dapat dilihat dari kenaikan nilai *precision*, *recall*, dan *f-Measure* pada ketiga dataset yang digunakan. Namun, tidak semua permasalahan *imbalanced* data dapat diselesaikan dengan teknik ini. Contoh kasus dapat dilihat pada *breast cancer* dataset dimana proses *resampling* menggunakan SMOTE-Tomek menurunkan *f-Measure*. Oleh karena itu, teknik ini harus digunakan dengan memperhatikan karakteristik data dan mesin klasifikasi yang digunakan.

5. DAFTAR PUSTAKA

[1] P. Branco., L. Torgo., dan R. Ribeiro. 2015. A Survey of Predictive Modelling under Imbalanced Distributions

[2] A. Hardoni., dan D. P. Rini. 2020. Integrasi Pendekatan Level Data Pada Logistic Regression untuk Prediksi Cacat Perangkat Lunak. *JIKO (Jurnal Informatika dan Komputer)*, 3 (2), pp. 101-106. DOI: 10.33387/jiko

[3] Chawla, N., et al. 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16. Hal. 321-357. doi: 10.1613/jair.953.

[4] H. Han., et al. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Advances in Intelligent Computing, Hal. 878-887. doi: 10.1007/11538059_91

[5] F. Last., et al. 2017. Oversampling for Imbalanced Learning Based on K-Means and SMOTE.

[6] C. Zhang., et al. 2018. A Cost-Sensitive Deep Belief Network for Imbalanced Classification.

[7] H. Haibo., et al. 2008. Adasyn: Adaptive synthetic samplingapproach for imbalanced learning. International Joint Conference on Neural Networks, June, 2008. 10.1109/IJCNN.2008.4633969.

[8] I. Tomek. 1976. Two modifications of CNN. Two Modifications of CNN. iIEEE Transactions on Systems, Man, and Cybernetics. Vol. SMC-6, No. 11: 769-772. doi: 10.1109/TSMC.1976.4309452.

[9] D. L. Wilson. 1972. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man, and Cybernetics. Vol. SMC-2, No. 3: 408-421, doi: 10.1109/TSMC.1972.4309137.

[10] M. Kubat dan S.Matwin. 1997. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. 14th International Conference on Machine Learning (ICML97) (USA: Tennessee) 179.

[11] J. Laurikkala. 2001. Improving Identification of Difficult Small Classes by Balancing Class Distribution. 8th Conference on AI in Medicine in Europe AIME01 (Portugal: Cascais) 63.

[12] N. Chamida., M. M. Santoni, dan N. Matondang. 2020. Pengaruh Oversampling pada Klasifikasi Hipertensi dengan Algoritma Naïve Bayes, Decision Tree, dan Artificial Neural Network (ANN). RESTI, 4 (4), pp. 635-641. https://doi.org/10.29207/resti.v4i4.2015.

[13] O. Heranova. 2019. Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring. RESTI, 3 (3), pp. 443-335. https://doi.org/10.29207/resti.v3i3.1275.

[14] R. A. Barro., I. D. Sulvianti., dan F. M. Afendi,. 2013. Penerapan Synthetic Minority Oversampling Technique (SMOTE) terhadap Data Tidak Seimbang pada Pembuatan Model Komposisi Jamu. Xplore, 1 (1), pp. 1-6. https://doi.org/10.29244/xplore.v1i1.12424

[15] S. Al-Azani dan E. El-Alfy. 2018. Imbalanced Sentiment Polarity Detection Using Emoji-Based Features and Bagging Ensemble. Pp. 1-5. 10.1109/CAIS.2018.8441956.

[16] C. Padurariu dan M. Breaban. 2019. Dealing with Data Imbalance in Text Classification. Procedia Computer Science. 159. 736-745. 10.1016/j.procs.2019.09.229.

[17] R. M. Pereira., Y. M. G. Costa dan C. N. Silla Jr. 2020. MLTL: A multi-label approach for the Tomek Link undersampling algorithm. Neurocomputing, 383, pp. 95-105. https://doi.org/10.1016/j.neucom.2019.11.076.

[18] R. M. Pereira. et. al. 2018. Dealing with Imbalanceness in Hierarchical Multi-Label Datasets Using Multi-Label Resampling Techniques. 818-824. 10.1109/ICTAI.2018.00128.

[19] A. Fernández., et al. 2018. Learning from Imbalanced Data Sets. 10.1007/978-3-319-98074-4.

[20] E. A. Sari., et al. 2020. Klasifikasi Kabupaten Tertinggal Di Kawasan Timur Indonesia dengan Support Vector Machine. *JIKO (Jurnal*

Informatika dan Komputer), 3 (3), pp. 188–195. DOI: 10.33387/jiko.v3i3.2364.

[21] L. Vig. 2014. Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset. Open Access Library Journal, 1, 1-7. doi: 10.4236/oalib.1100660.

[22] H. Asri. et al. 2016. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science. 83. 1064-1069. 10.1016/j.procs.2016.04.224.

[23] S. Ubaidillah., R. Sallehuddin., dan N. A. Ali, N. 2013. Cancer Detection Using Aritifical Neural Network and Support Vector Machine: A Comparative Study. Jurnal Teknologi. 65. 10.11113/jt.v65.1788.

[24] A. Kabir., S. Basuki dan G. W. Wicaksono. 2019. Analisis sentimen kritik dan saran pelatihan aplikasi teknologi informasi (PATI) menggunakan algoritma support vector machine (SVM). Repositor. 1. 10.22219/repositor.v1i1.11.

[25] P. Sinha., dan P. Sinha. 2015. Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM. International Journal of Engineering Research and. V4. 10.17577/IJERTV4IS120622.