

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini memiliki tujuan untuk menangani distribusi kelas tidak seimbang pada dataset menggunakan teknik *resampling oversampling* ADASYN k-NN dan reduksi fitur pada dataset menggunakan jenis fitur seleksi filter yaitu Information Gain. Berdasarkan pada penelitian yang telah dilakukan dapat ditarik kesimpulan sebagai berikut:

1. Algoritma *Decision Tree* implementasi *resampling* ADASYN dan fitur seleksi *Information Gain* mampu meningkatkan performa hasil *geometric-mean* baik dalam pelatihan maupun pengujian.
2. Algoritma *Decision Tree* dengan penanganan ketidakseimbangan kelas dengan *resampling* ADASYN mampu memperbaiki kemampuan klasifikasi dalam pelatihan maupun pengujian dan penggunaan fitur seleksi berpotensi menurunkan nilai hasil evaluasi untuk *dataset libras move, mammography, wine quality*, dan *yeast ml8*, berbanding terbalik implementasi fitur seleksi dalam dataset *ecoli* mampu meningkatkan nilai akurasi dan *geometric mean*, kondisi ini bergantung pada karakteristik data dan keberadaan *noise sample* setiap dalam dataset.
3. Berdasarkan hasil dari hasil uji akurasi dan uji *geometric mean* algoritma *Naïve Bayes* mencapai performa terbaik pada skenario kedua, yaitu *resampling* tanpa fitur seleksi.

4. Implementasi resampling ADASYN pada Algoritma *Naïve Bayes* dapat meningkatkan nilai *geometric mean* dan fitur seleksi memiliki kecenderungan menurunkan performa dari nilai *geometric mean*.
5. Dataset *Ecoli* dengan *imabalance rasio* 8.6 yang diklasifikasikan dengan algoritma *Naïve Bayes* memiliki kemampuan terbaik pada skenario ke-3 dengan hasil akurasi=90% dan nilai *geometric mean*=89% dimana hasil *train* dan *test* memiliki perubahan rentang nilai rendah menandakan model mencapai kondisi optimal dan tidak terdapat indikasi terjadinya *overfitting*.
6. Algoritma *Decision Tree* mencapai nilai akurasi dan *geometric mean* terbaik pada *dataset* *libras move imbalance ratio* (IR) 14, akurasi = 99.25%, *geometric mean*= 99.31%.
7. Dataset *mammography* memiliki *imbalance rasio* (42) lebih optimal diklasifikasi dengan menggunakan algoritma *Decision Tree* dengan perlakuan penanganan terhadap *imbalance class* sehingga di dapat nilai *geometric mean* = 94,92% dan nilai akurasi = 94,98%.
8. Penanganan terhadap *dataset imbalance wine quality* dengan *imbalance rasio* 26 dan *dataset yeast m18* dengan *imbalance rasio* 13 memberikan hasil evaluasi performa lebih tinggi pada algoritma *Decision Tree* namun algoritma *Naïve Bayes* memberikan hasil evaluasi yang lebih stabil untuk setiap skenario yang diberikan.
9. Dataset *mammography* memiliki karakteristik ketidakseimbangan kelas yang baik menurut distribusi data, dataset ini mampu diklasifikasi oleh

algoritma klasifikasi *Naïve Bayes* dan *Decision Tree* dengan baik walaupun tanpa adanya perlakuan penanganan ketidakseimbangan distribusi pada kelas *dataset*.

10. Penanganan *dataset imbalance* menggunakan metode resampling *oversampling* ADASYN dalam penelitian ini tidak selalu memberikan peningkatan nilai akurasi, namun terjadi perbaikan nilai *geometric mean* baik dalam pelatihan maupun pengujian, sehingga kelas minoritas lebih mudah dikenali walaupun dengan nilai akurasi lebih rendah.

5.2 Saran

Berdasarkan pada penelitian yang sudah dilakukan terdapat saran untuk penelitian selanjutnya terkait dengan masalah dataset imbalance:

1. Penelitian selanjutnya dapat diujikan dengan menggunakan dataset *multiclass*.
2. Penelitian selanjutnya dapat digunakan dataset sejenis dengan variasi *imbalance ratio*.
3. Dataset yang digunakan dapat diujikan pada metode *undersampling* atau kombinasi (*oversampling-undersampling*).
4. Pada penelitian selanjutnya dapat dilakukan penanganan baik level data menggunakan *resampling* dikombinasikan dengan algoritma *ensemble*.
5. Dapat digunakan fitur seleksi selain *information gain*, seperti CBFS, Mutual Information, atau jenis fitur seleksi selain metode filter, seperti *wrapper* atau *embedded*.

6. Dapat dilakukan pengujian perbedaan pemilihan nilai beta terhadap hasil sintesis.
7. Pengujian nilai optimal untuk ambang batas bawah fitur tetap dipertahankan.

DAFTAR PUSTAKA

- [1] C. M. Bishop, "Bishop - Pattern Recognition and Machine Learning - Springer 2006," *Antimicrob. Agents Chemother.*, 2014, doi: 10.1128/AAC.03728-14.
- [2] S. Russell and P. Norvig, *Artificial Intelligence A Modern Approach Third Edition*. 2010.
- [3] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning (Adaptive Computation and Machine Learning series)*. 2012.
- [4] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, 2009, doi: 10.1142/S0218001409007326.
- [5] M. Bach, A. Werner, J. Żywiec, and W. Pluskiewicz, "The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis," *Inf. Sci. (Ny)*, 2017, doi: 10.1016/j.ins.2016.09.038.
- [6] Y. Pristyanto, S. Adi, and A. Sunyoto, "The effect of feature selection on classification algorithms in credit approval," *2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019*, pp. 451-456, 2019, doi: 10.1109/ICOIACT46704.2019.8938523.