JT-IBSI (Jurnal Teknik Ibnu Sina), Vol 9, No. 01, Juni 2024
E-ISSN: 2541-2647 DOI: https://doi.org/10.36352/jt-ibsi.v9i01.807

Penyelesaian Masalah Ketidakseimbangan Data Melalui Teknik *Oversampling* dan *Undersampling* pada Klasifikasi Siswa Tidak Naik Kelas

[1)]Moch. Anjas Aprihartha, [2)]Dicky Zulhan, [3)]Fatma Ahardika Nurfaizal, [4)]Taufik Nur Alam
[1,2,3,4)] Program Studi PJJ Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
e-mail: anjas.aprihartha@dsn.dinus.ac.id

Abstract

Data mining is the process of generating patterns and knowledge from large datasets. Data sources can be obtained from databases, the web, or other information storage. Most data mining algorithms work best when the number of samples in each class is almost the same. But in the case of classification problems, the number of observations belonging to one class is significantly smaller than that of other classes is not a rare thing at all. This is called imbalanced data. To overcome the problem of data imbalance, resampling techniques can be used. Resampling is divided into two types, namely undersampling and oversampling. This research will apply oversampling and undesampling techniques followed by classification predictions using the C5.0 algorithm in the case of classification of students who do not graduate from school. Based on the test results that have been carried out with three different datasets, the C5.0 algorithm with k-fold cross validation can work better on datasets processed using random oversampling techniques compared to original datasets or datasets formed from random undersampling techniques. This is indicated by the accuracy in each fold which tends to be stable and consistent in the range of 93% to 97.6%.

***Keywords**:* C5.0, *data mining, k-fold cross validation, oversampling, undersampling*

Abstrak

Data mining adalah proses menghasilkan pola dan pengetahuan dari himpunan data besar. Sumber data dapat diperoleh dari database, web, atau penyimpanan informasi lainnya. Sebagian besar algoritma data mining bekerja paling baik ketika jumlah sampel di setiap kelas berada pada jumlah yang hampir sama. Namun dalam kasus masalah klasifikasi, jumlah observasi yang termasuk dalam satu kelas sangatlah signifikan lebih kecil daripada kelas lain bukanlah hal yang langka sama sekali. Inilah yang disebut klasifikasi data yang tidak seimbang. Dalam mengatasi masalah ketidakseimbangan data dapat dilakukan teknik *resampling*. Pengambilan sampel ulang dibagi menjadi dua jenis, yaitu pengambilan sampel terlalu rendah yang disebut *random undersampling* dan pengambilan sampel berlebihan yang disebut *random oversampling*. Penelitian ini akan menerapkan teknik *oversampling* dan *undesampling* yang dilanjutkan dengan prediksi klasasifikasi menggunakan algoritma C5.0 pada kasus klasifikasi siswa yang tidak naik kelas disekolah. Berdasarkan hasil uji yang telah dilakukan dengan tiga dataset yang berbeda, algoritma C5.0 dengan k-*fold cross validation* dapat bekerja lebih baik pada dataset yang diolah melalui teknik *random oversampling* dibandingkan dataset asli ataupun dataset yang dibentuk dari teknik *random undersampling*. Ini ditunjukkan dengan akurasi pada setiap lipatan cenderung stabil dan konsisten pada kisaran 93% sampai 97,6%.

Kata kunci: C5.0, *data mining, k-fold cross validation, oversampling, undersampling*

Diterima : Juni 2024
Disetujui : Juni 2024
Dipublikasi : Juni 2024

Pendahuluan

Data mining adalah proses menghasilkan pola dan pengetahuan dari himpunan data besar. Sumber data dapat diperoleh dari database, web, atau penyimpanan informasi lainnya (Han *et. al*, 2012). Data mining dapat mengatasi kumpulan data yang besar secara terbuka, sehingga tidak mungkin untuk memberikan batasan ketat pada pertanyaan yang ingin diselesaikan yang memerlukan inferensi (Shmueli, 2017). Akibatnya, pendekatan umum terhadap *data mining* rentan terhadap bahaya *overfitting*, yaitu ketika suatu model sangat cocok dengan sampel data yang tersedia sehingga model tersebut tidak hanya menggambarkan karakteristik struktural data. Hal ini dapat terjadi apabila pada data memiliki jumlah kelas dengan perbedaan sangat signifikan.

Sebagian besar algoritma *data mining* bekerja paling baik ketika jumlah sampel di setiap kelas berada pada jumlah yang hampir sama. Namun dalam kasus masalah klasifikasi, jumlah observasi yang termasuk dalam satu kelas sangatlah signifikan lebih kecil daripada kelas lain bukanlah hal yang langka sama sekali. Inilah yang disebut klasifikasi data yang tidak seimbang (Kantardzic, 2011). Dalam penerapan praktis, rasio kelas kecil dan kelas besar bisa sangat drastis sebesar 1:10, sementara beberapa aplikasi untuk deteksi penipuan melaporkan ketidakseimbangan sebesar 1:100.000. Kasus serupa terjadi di bidang pendidikan, database catatan yang dapat digunakan untuk membangun model prediksi klasifikasi untuk siswa yang tidak naik kelas.

Pengambilan sampel ulang *(resampling)* sebagai pendekatan yang efektif untuk menangani kumpulan data yang tidak seimbang, yang bertujuan untuk menyamakan jumlah sampel kategori (Guan, *et. al*., 2024). Pengambilan sampel ulang sebagian besar mencakup dua metode, yaitu pengambilan sampel terlalu rendah yang disebut *random undersampling* dan pengambilan sampel berlebihan yang disebut *random oversampling.*

Telah banyak penelitian yang dilakukan dalam mengatasi masalah ketidakseimbangan data. Penelitian oleh Prasetya (2022), menerapkan teknik *oversampling* dan *undersampling* pada kasus kanker serviks pada implementasi naive bayes. Hasil penelitian diperoleh teknik *undersampling* memberikan akurasi yang lebih baik dibandingkan teknik *oversampling*. Penelitian oleh Abdani *et. al* (2022) yang meneliti ketidakseimbangan data dalam mendeteksi sel kanker untuk leukemia limfoblas akut. Hasil uji menunjukkan bahwa teknik *oversampling* memberikan peningkatan akurasi 0,7754 menjadi 0,7807.

Berdasarkan paparan yang telah dijelaskan, studi ini akan menerapkan teknik *oversampling* dan *undersampling* yang dilanjutkan dengan menguji masing-masing dataset dengan algoritma C5.0 untuk mengetahui performa masing-masing model pada dataset yang berbeda. Kasus yang menjadi objek penelitian ini diperoleh dari Setiawan (2020) yang meneliti tentang klasifikasi kenaikan kelas di SDN Citamiang 2 dengan algoritma C4.5. Pada data yang dikumpulkan

terindikasi kelas yang tidak seimbang, sehingga menarik perhatian penulis dalam membuat paper ini.

Metode Penelitian

1. Data dan Variabel Penelitian

Data dalam penelitian ini merupakan data sekunder dari Setiawan (2020). Pada dataset terdiri dari 239 amatan yang mengandung kelas tidak seimbang. Variabel dan jenis data disajikan pada Tabel 1.

Tabel 1. Variabel dan Jenis Data

Variabel	Jenis Data
Jenis Kelamin	Kategorik
Nilai Afektif	Numerik
Kehadiran	Numerik
Jumlah Maple Her	Kategorik
Kelas	Kategorik

2. *Bootstrap*

Metode ini mengambil sampel ulang data yang tersedia dengan penggantian untuk menghasilkan sejumlah kumpulan data "palsu" dengan ukuran yang sama pada kumpulan data yang diberikan (Kantardzic, 2011). Jumlah dataset ini biasanya mencapai beberapa ratus. Kumpulan pelatihan baru ini dapat digunakan untuk menyelesaikan masalah keseimbangan data. Metode ini sangat berguna dalam situasi kumpulan data kecil.

Pada proses *bootstrap*, setiap data mempunyai probabilitas terambil sebesar $1/d$ dan probabilitas tidak terambil sebesar $1 - 1/d$. Jika diambil sebanyak *d* kali maka probabilitas yang tidak terambil pada saat tersebut adalah $\left(1 - 1/d\right)^d$ (Han *et. al*, 2011).

$$Q = \underbrace{\left(1 - 1/d\right)\left(1 - 1/d\right)\cdots\left(1 - 1/d\right)}_{d} = \left(1 - 1/d\right)^d \quad (1)$$

3. Algoritma C5.0

Terdapat beberapa penerapan dari algoritma pohon keputusan, salah satu yang paling baik adalah yang dikenal adalah algoritma C5.0 (Lantz, 2019). Algoritma ini merupakan versi perbaikan dari algoritma sebelumnya, yaitu C4.5, yang merupakan algoritma versi peningkatan dari algoritma Iterative Dichotomizer 3 (ID3) (Rajeswari & Suthendran, 2019).

Algoritma C5.0 menggunakan *entropy* sebagai ukuran kemurnian yang digunakan untuk mengidentifikasi kandidat pemisahan pohon keputusan yang terbaik. Biasanya, *entropy* diukur dalam bit. Jika hanya ada dua kelas yang mungkin, *entropy* nilai dapat berkisar dari 0 hingga 1. Untuk n kelas, *entropy* berkisar dari 0 hingga $\log_2(n)$. Misalkan himpunan data tertentu (S), istilah c mengacu pada jumlah tingkat kelas, dan p_i mengacu pada proporsi nilai yang termasuk dalam tingkat kelas i. Dalam pemahaman matematika *entropy* didefisikan sebagai (Lantz, 2019).

$$entropy(S) = \sum_{i=1}^{c} -p_i \log_2(p_i) \quad (2)$$

Untuk menggunakan *entropy* guna menentukan fitur optimal untuk dipecah, algoritma menghitung perubahan homogenitas yang akan dihasilkan dari pemisahan pada setiap kemungkinan fitur, ukuran yang dikenal sebagai *information gain*. Perolehan informasi untuk fitur F dihitung sebagai selisih antara *entropy* pada segmen sebelum pemisahan (S_1) dan partisi hasil split (S_2).

$$InfoGain(F) = entropy\,(S_1) - entropy\,(S_2) \quad (3)$$

Dalam perhitungannya $entropy\,(S_2)$ mempertimbangkan total *entropy* disemua partisi. Hal ini dilakukan dengan memberi bobot pada *entropy* setiap partisi sesuai dengan proporsi yang telah dihitung pada partisi tersebut.

$$entropy(S) = \sum_{i=1}^{n} w_i\, entropy(p_i) \quad (4)$$

Total *entropy* yang dihasilkan dari pemisahan adalah jumlah *entropy* dari masing-masing *n* partisi yang diberi bobot berdasarkan proporsi contoh yang termasuk dalam partisi (w_i).

4. *K-Fold Cross Validation*

K-fold cross validation merupakan teknik yang secara acak membagi data menjadi *k* bagian yang berbeda. Kemampuan algoritma menjadi sensitif terhadap *k* apabila nilai *k* yang ditentukan sangat kecil (Chacón *et al.*, 2023). Umumnya nilai k yang ditetapkan adalah 5 atau 10. Tahapan dalam melakukan *k-fold cross validation sebagai* berikut.

1. Data pelatihan dipartisi menjadi *k* lipatan.
2. Proses pelatihan dan validasi dijalankan berulang sebanyak *k* kali.
    a. Salah satu lipatan dimanfaatkan sebagai validasi.
    b. Sisanya digunakan untuk melatih algoritma.
    c. Algoritma diuji pada lipatan validasi dengan perhitungan akurasi.
3. Hasil kinerja algoritma dari setiap perulangan dihitung rata-ratanya untuk mendapatkan performa model secara simultan.

5. Akurasi

Untuk mengetahui tingkat keberhasilan suatu algoritma klasifikasi maka dapat diukur dengan melihat seberapa sukses algoritma dapat memutuskan data yang masuk pada kelas yang tepat. Tingkat keberhasilan atau akurasi dapat dirumuskan sebagai berikut (Ramasubramanian & Singh, 2017).

$$akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Keterangan:

• *True Positive* (TP): Diklasifikasikan dengan benar sebagai kelas positif
• *True Negative* (TN): Diklasifikasikan dengan benar sebagai kelas negatif
• *False Positive* (FP): Diklasifikasikan dengan salah sebagai kelas positif
• *False Negative* (FN): Diklasifikasikan dengan salah sebagai kelas negatif

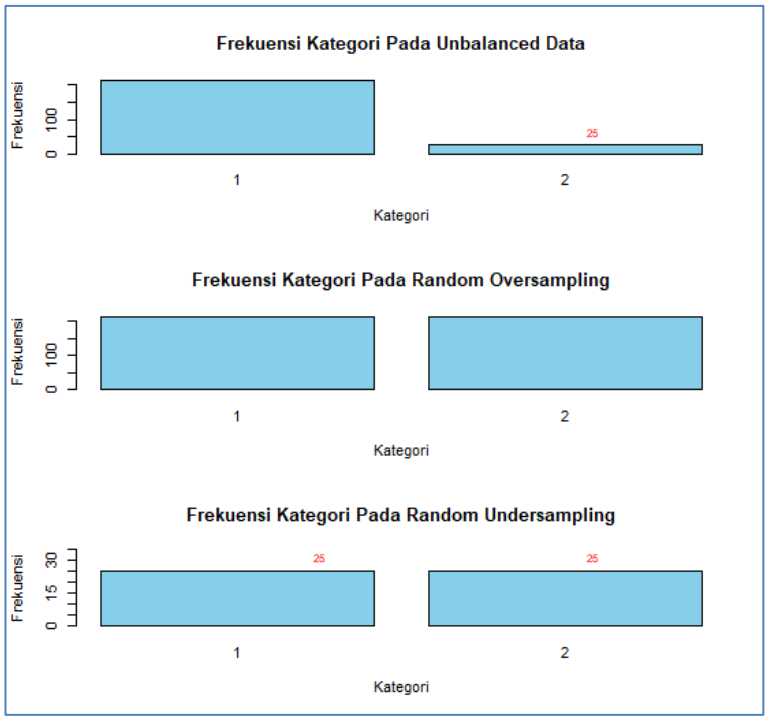
Hasil dan Pembahasan

1. Algoritma *Random Oversampling* dan *Random Undersampling*

Pada Gambar 1 menampilkan diagram batang pada frekuensi kategori yang berbeda-beda. Diagram batang pertama menunjukkan frekuensi data asli dengan frekuensi kategori naik kelas (kelas 1) jauh lebih tinggi dibandingkan kategori tidak naik kelas (kelas 2). Jumlah sampel yang berada pada kelas 1 sebanyak 214 amatan sedangkan pada kelas 2 sebanyak 25 amatan. Oleh karena dataset terdeteksi adanya ketidakseimbangan data maka perlu dilakukan manipulasi data dengan algoritma *random undersampling* atau *random oversampling*.

Diagram batang kedua menunjukkan hasil menggandakan data pada kelas minoritas, yaitu kelas 2 dengan algoritma *random oversampling*. Proses duplikasi melalui pengambilan data acak satu persatu dengan pengembalian sebanyak 189 kali dari kumpulan data pada kelas 2. Diperlihatkan bahwa total amatan pada kelas 2 bertambah menjadi 214 amatan, sehingga total sampel menjadi 428 amatan dengan jumlah amatan masing-masing kelas sebanyak 214 amatan.

Diagram batang ketiga menunjukkan hasil pemotongan data pada kelas mayoritas, yaitu kelas 1 dengan algoritma *random undersampling*. Proses reduksi dilakukan dengan mengambil data acak satu persatu dari kelas 1 sebanyak 25 kali dengan pengembalian. Semua data yang telah terambil kemudian dikumpulkan, lalu digabungkan dengan kumpulan data pada kelas 2. Total sampel menjadi 50 amatan dengan jumlah amatan masing-masing kelas sebanyak 25 amatan.



Gambar 1. Frekuensi Ketiga Dataset

2. Perbandingan Model C5.0 Terhadap Keseimbangan Data

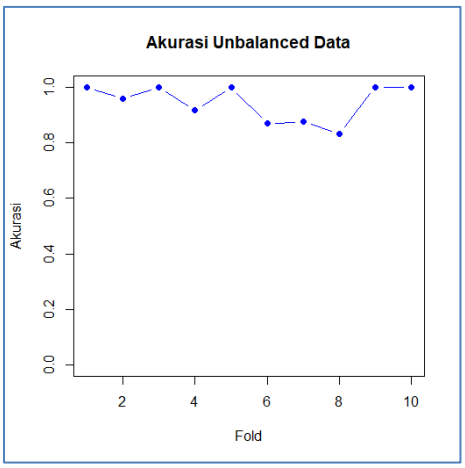
Penelitian ini menggunakan tiga jenis dataset. Dataset pertama adalah data asli dengan kelas yang tidak seimbang, dataset kedua dan ketiga masing-masing merupakan data hasil proses *random oversampling* dan *random undersampling.* Ketiga dataset akan diuji dengan algoritma C5.0 untuk melihat perbedaan masing-masing performa model ketiga dataset. Untuk memaksimalkan pemanfaatan dataset maka diterapkan teknik *k-fold cross validation* pada algoritma C5.0. Tujuannya agar semua data memiliki peluang yang sama menjadi dataset *training* dan dataset *testing* sehingga algoritma dapat mengevaluasi secara simultan pada keseluruhan dataset. Jumlah lipatan yang ditetapkan sebanyak 10 lipatan.

Gambar 2 menunjukkan plot garis akurasi sepuluh lipatan pada dataset tidak seimbang. Terlihat bahwa akurasi berada pada rentang 0,833 (83,3%) sampai 1 (100%). Beberapa lipatan mencapai akurasi 100%. Rata-rata akurasi yang dihasilkan sebesar 0,945 (94,5%) yang mengindikasikan model sangat baik dalam klasifikasi data, namun akurasi setiap lipatan cenderung tidak stabil. Hal ini disebabkan algoritma kesulitan mengenali data yang tidak seimbang. Algoritma cenderung memprediksi data pada kelas mayoritas (kelas 1) dibandingkan kelas minoritas (kelas 2).

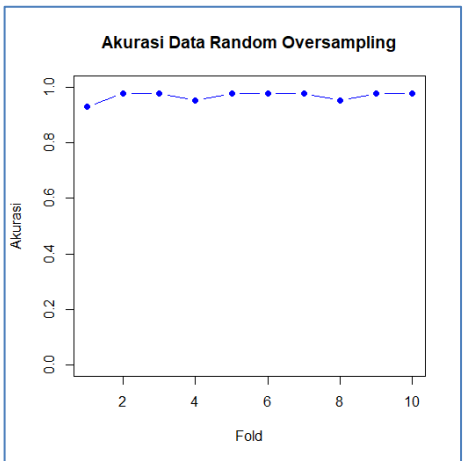
Gambar 3 menunjukkan plot garis pada dataset *random oversampling*. Terlihat bahwa akurasi setiap lipatan berada pada rentang 0,93 (93%) sampai 0,976 (97,6%). Rata-rata akurasi yang dihasilkan sebesar 0,967 (96,7%) yang berarti model sangat baik dalam klasifikasi data. Plot

memperlihatkan variasi akurasi sangat stabil dan kosisten. Hal ini disebabkan, algoritma dapat mengenali model dengan baik pada setiap lipatan sehingga menghasilkan akurasi yang sangat baik.

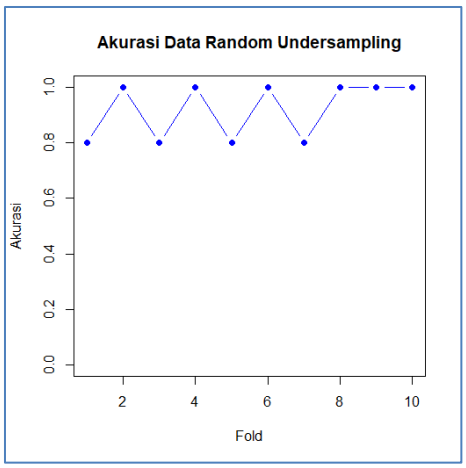
Gambar 4 menunjukkan plot garis akurasi dataset *random undersampling*. Akurasi berada pada rentang 0,8 (80%) hingga 1 (100%). Rata-rata akurasi yang dihasilkan sebesar 0,92 (92%). Plot memperlihatkan variasi akurasi yang tidak konsisten, terdapat pola zigzag secara bergantian antara 0,8 dan 1. Hal ini disebabkan algoritma kesulitan memprediksi data karena hilangnya informasi penting dalam dataset, sehingga memberikan kinerja model yang berbeda disetiap lipatannya.



Gambar 2. Akurasi Dataset Asli



Gambar 3. Akurasi Dataset *Oversampling*



Gambar 4. Akurasi Dataset *Undersampling*

Pembahasan

Berdasarkan hasil uji yang telah dilakukan dengan tiga dataset yang berbeda diperoleh bahwa algoritma C5.0 dapat bekerja lebih baik pada dataset yang diolah melalui teknik *random*

oversampling dibandingkan dataset asli ataupun dataset yang dibentuk dari teknik *random undersampling*. Ini ditunjukkan dengan akurasi pada setiap lipatan cenderung stabil dan konsisten pada kisaran 93% sampai 97,6%.

Menurut Li (2024), pada tingkat data, metode *oversampling* umumnya lebih efektif dibandingkan metode *undersampling* maupun *hybrid*. Ini disebabkan proses *undersampling* atau *hybrid* dilakukan pembuangan sampel yang berisi informasi penting. Teknik *oversampling* menawarkan keuntungan dalam meningkatkan fokus klasifikasi pada kelas minoritas dan meningkatkan kinerjanya dalam mengklasifikasikan sampel kelas minoritas (Wang, *et al*., 2024). Teknik ini layak digunakan untuk skenario apabila terdapat disparitas data yang signifikan antara sampel masuk kelas minoritas dan mayoritas. Ini dikarenakan pengambilan sampel yang berlebihan memungkinkan retensi informasi penting dari data asli (Wang, *et al*., 2024; Brieman, 2021). Hal ini membantu dalam mitigasi kehilangan informasi yang berlebihan dan mengurangi masalah kehilangan informasi yang dapat terjadi akibat *undersampling*.

Kesimpulan

Pada penelitian dapat disimpulkan bahwa data asli mengandung kelas tidak seimbang diuji dengan algoritma C5.0 dan 10-*fold cross validation*, diperoleh akurasi berada pada rentang 83,3% sampai 100%. Rata-rata akurasi yang dihasilkan sebesar 94,5% yang mengindikasikan model sangat baik dalam klasifikasi data, namun akurasi setiap lipatan cenderung tidak stabil. Pada metode random *oversampling*, amatan bertambah pada kelas minoritas sebanyak 214 amatan, sehingga total data keseluruhan menjadi 428 amatan. Hasil uji diperoleh akurasi setiap lipatan berada pada rentang 93% sampai 97,6%. Rata-rata akurasi yang dihasilkan sebesar 96,7% yang berarti model sangat baik dalam klasifikasi data. Pada metode random *undersamping*, amatan tereduksi pada kelas mayoritas menjadi 25 amatan, sehingga total keseluruhan data menjadi 50 amatan. Akurasi berada pada rentang 80% hingga 100%. Rata-rata akurasi yang dihasilkan sebesar 0,92 (92%). Berdasarkan rata-rata akurasi dari tiga dataset yang berbeda, algoritma C5.0 dapat bekerja lebih baik pada dataset *oversampling* dibandingkan dataset asli ataupun dataset *undersampling*.

Daftar Pustaka

Abdani, S. R., Zulkifley, M. A., & Zulkifley, N. H. (2022, March). Undersampling and oversampling strategies for convolutional neural networks classifier. In *Proceedings of the 6th International Conference on Electrical, Control and Computer Engineering: Inecce2021, Kuantan, Pahang, Malaysia, 23rd August* (pp. 1129-1137). Singapore: Springer Singapore.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Chacón, A. M. P., Ramírez, I. S., & Márquez, F. P. G. (2023). K-nearest neighbour and K-fold cross-validation used in wind turbines for false alarm detection. *Sustainable Futures*, *6*, 100132.

Guan, S., Zhao, X., Xue, Y., & Pan, H. (2024). Awgan: An adaptive weighting Gan approach for oversampling imbalanced datasets. *Information Sciences*, 120311.

Han, J., Kamber, M., & Pei, J. (2012). Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*.

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.

Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.

Li, J. (2024). Oversampling Framework Based on Sample Subspace Optimization with Accelerated Binary Particle Swarm Optimization for Imbalanced Classification. *Applied Soft Computing*, 111708.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). *Data mining for business analytics: concepts, techniques, and applications in R*. John Wiley & Sons.

Prasetya, J. (2022). Penerapan Klasifikasi Naive Bayes dengan Algoritma Random Oversampling dan Random Undersampling pada Data Tidak Seimbang Cervical Cancer Risk Factors. *Leibniz: Jurnal Matematika*, *2*(2), 11-22.

Rajeswari, S., & Suthendran, K. (2019). C5. 0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud. *Computers and Electronics in Agriculture*, *156*, 530-539.

Ramasubramanian, K., & Singh, A. (2017). *Machine learning using R* (No. 1). New Delhi, India: Apress.

Setiawan, A. (2020). Penerapan algoritma C.45 untuk klasifikasi tingkat kenaikan kelas di Sdn Citamiang 2 (Skripsi, Universitas Bina Sarana Informatika). Universitas Bina Sarana Informatika.

Wang, F., Zheng, M., Hu, X., Li, H., Wang, T., & Chen, F. (2024). Fiao: Feature Information Aggregation Oversampling for Imbalanced Data Classification. *Applied Soft Computing*, 111774.